A BIOINFORMATIC STUDY OF SEPTIN EVOLUTION AND CO-EVOLUTION

by

FANGFANG PAN

(Under the Direction of Michelle Momany and Russell L. Malmberg)

ABSTRACT

Septins are filamentous GTP-binding proteins that have been found in animals, fungi and microsporidia. This dissertation focuses on a computational study of the septin gene family. The phylogenetic analysis revealed orthologous relationships of septins across kingdoms. It divided all septins into five groups and suggested a consistent nomenclature. Septins interact with themselves as well as other proteins. In protein co-evolution, the interacting proteins change but still maintain protein-protein interactions. A computational method using mutual information was developed to study protein co-evolution and predict protein residue physical interactions for co-evolving positions. This method was applied to the septin gene family to identify co-evolution and interactions among septin subunits, as well as between septins and two proteins, formin and myosin.

INDEX WORDS: Septin, Evolution, Protein Co-evolution, Protein interaction

A BIOINFORMATIC STUDY OF SEPTIN EVOLUTION AND CO-EVOLUTION

by

FANGFANG PAN

B.A., Fudan University, China, 2002

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2007

A BIOINFORMATIC STUDY OF SEPTIN EVOLUTION AND CO-EVOLUTION

by

FANGFANG PAN

Major Professor:	Michelle Momany Russell L. Malmberg
Committee:	Claiborne V. C. Glover III Liming Cai Jaxk Reeves

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
December 2007

DEDICATION

To my family

ACKNOWLEDGEMENTS

I would like to acknowledge my gratitude to my major professors Dr. Michelle Momany and Dr. Russell Malmberg for their patience, encouragement and guidance. These studies could not have been done without their help.

I wish to thank my committee members Dr. Claiborne Glover, Dr. Liming Cai and Dr. Jaxk Reeves for their support.

I acknowledge the assistance from former and present members of Momany, Malmberg and Cai labs.

Special thanks to Dongsheng Che, who helped me through a time when no one else could have.

TABLE OF CONTENTS

Page

CHAPTER 1

INTRODUCTION

The fast growing quantities of molecular data have led to the development of bioinformatics [1]. As a collaborative discipline integrating computer science with biology, bioinformatics provides a variety of application tools for organizing and analyzing biological data. These vary from programs to analyze DNA sequences and gene expressions to protein structure predictions and protein interaction predictions. The analyses are on a large genome scale as well as are focused on specific genes. The work described here includes two major fields of bioinformatics, phylogenetic analysis and protein-protein interactions, and is focused on and applied to the septin gene family.

The septin gene family

Septin proteins were first found at the mother-bud neck of budding yeast *Saccharomyces cerevisiae,* where they are associated with a ring structure in the plasma membrane around the neck and are important for cytokinesis [2]. Later, septins were found in many other fungi in addition to *S. cerevisiae,* as well as in animals. Currently, septins have been identified in more than 20 organisms, including fungi, animals, and microsporidia [3]. No septins have been found in the plant kingdom. In animals, septins exist in a wide range of organisms, from sponges and nematodes, to fish and mammals. In fungi, septins were found in several different lineages, including chytridiomycetes, basidiomycetes and ascomycetes. Microsporidia has been regarded as a sister group of fungi or as extremely reduced fungi. Three septins were found in *Encephalitozoon cuniculi,* the first microsporidium that has had its whole genome sequenced. With the fast growth of genome sequencing projects, more septins are rapidly being identified.

Within each organism, the number of genes that belong to the septin family also varies a lot. In the organisms that have complete genomes, the number of identified septins can range from two in *Caenorhabditis elegans* to thirteen in *Homo sapiens*. In *S. cerevisiae,* where the septins were first identified, there are seven septins genes, Cdc3, Cdc10, Cdc11, Cdc12, Shs1, Spr3, Spr28. The protein products of the first five septin genes form a heteropentamer complex and assemble into the filaments of 10nm diameter [4].

Together with associated proteins, septins form a collar structure during cell division along the axis of cell splitting [2, 5]. The filamentous septin ring structure has been found in a variety of places, such as fungal spores, germ tubes, bud neck, hyphae of filamentous fungi, pseudohyphal projections, mating projections, and the cell division plane of animal cells [6-9].

Septins exist in a wide range of organisms. Their functions show both consistency and variability in different organisms. New functional roles are still being discovered. Some functions are connected and can't completely be separated from one another. In some fungi, septins are required for bud or hyphae site selection [10, 11]. Another important septin function is to form a diffusion barrier at the neck, restrict the motion of plasma membrane proteins and divide cells into different compartments for polarity [12, 13]. Septins can also work as a scaffold to recruit proteins in the yeast morphogenesis check point and during sporulation [6, 13, 14]. Similar functions were also found in animal cell cytokinesis [15, 16]. In animals, septins show some additional functions, such as vesicle trafficking, tumorigenesis, apoptosis, and cell movement [17, 18]. Though septins have been studied for around 30 years, some discoveries are very new and there are still many aspects to explore.

As proteins that are typically around 500 amino acids long, septins have some amino acid regions that are more conserved than others (Figure 1.1). Septins are P-loop GTPase proteins and

they are characterized by a GTPase domain, which contains three GTPase motifs, G1, G3 and G4 [7, 19]. Septins belong to the TRAFAC class of P-loop super class [19]. More specifically, septins belong to the septin-like family, which includes septins, paraseptins and related GTPases. Efforts have been put into finding the origin of septins. The sequences of the septin gene family were found to be more similar to some bacterial GTPases than to other TRAFAC class members [19]. Leipe *et al.* suggested that the septin gene family arose by horizontal transfer from bacterial GTPases to eukaryotes and the patchy distribution in eukaryotes resulted from lineage-specific gene loss. If septins could be found in some lower eukaryotic organisms, it might suggest alternative explanations of septin origin. In addition to the conserved GTPase domain, there is also a septin unique domain right after the GTPase domain [4]. There are variable regions on the two ends of the sequence. At the C-terminus, there is a coiled-coil region which is involved in protein-protein interactions [20]. At the N-terminus, septins have a variable region and a polybasic region before the GTPase domain.

Phylogenetic analysis and sequence conservation

Evolution is a branching process starting from one origin. Over time, populations may become extinct, change and divide into different branches. Phylogenetics is the study of evolution. It tries to reconstruct the lineage history by using tree diagrams to represent the relationships of different species.

There are various computational methods to reconstruct evolutionary history. “Distance methods” calculate the overall similarity of sequences by using matrices containing pair wise distance values and then clustering the closest. “Character-based methods" include parsimony, maximum likelihood and bayesian inference methods. They carry out calculations on each of the individual residues. Parsimony method builds trees for each nucleotide position and selects the

shortest tree that contains smallest number of overall changes. Maximum likelihood method uses explicit sequence evolutionary models and finds a tree that has the highest probability to give rise to the observed data. Distance and parsimony methods are usually very fast but have many limitations. The maximum likelihood method has lower variances and is more robust to violations, but it is computationally intensive.

Bayesian inference produces a posterior probability distribution of trees by using Markov Chain Monte Carlo (MCMC) simulation, based on the prior probability and the observed data [21, 22]. Through the implementation of the MCMC algorithm, Bayesian inference increases the analysis speed, and allows more extensive searches than previously possible, although it is still computationally intensive [23]. It is preferable for large datasets.

Conserved regions of a sequence are usually of interest to people as they may imply important sites that are kept during evolution. There are many computational methods for identifying conserved motifs. Entropy, which measures the uncertainty associated with a random variable, is a simple but effective statistic [24].

Protein co-evolution

Mutations can create changes in protein sequences. Some mutations cause the loss of essential protein functions that are required by the organism to survive. Natural selection removes less favorable mutations, and accumulates advantageous ones.

Many proteins work as units in a large interaction network. If the interactions between proteins are abolished by a mutation and the proteins fail to execute required functions, the mutations are deleterious. However, if compensatory mutations happen elsewhere which can restore the protein-protein interaction, that mutation is no longer selected against and thus two mutations

may be kept together during evolution. This type of correlated mutation process is called co-evolution.

These co-evolving mutations are important but are often neglected. At the amino acid sequence level, we may find low levels of conservation if we look at only one column in a multiple alignment. However, those sites might be very important to proteins. They may be crucial for protein tertiary structure, stabilization and interaction.

There have been efforts to detect correlated mutations computationally and experimentally [25-29]. Mutual information (MI) has been used to identify sites of correlated mutations in homologous sequence alignments [30]. Though MI is insensitive, requiring many sequences, the large number of currently available homologous sequences for many proteins makes it a good method to explore protein co-evolution. Other methods that have been applied to explore protein co-evolution include correlation statistics, phylogenetic trees, and bayesian and maximum likelihood estimation techniques [25, 26, 31-36]

Co-evolution of amino acid residues may help to predict protein 3D structures. Compensating mutations may occur when residues, which interact closely in 3D distance, mutate simultaneously and still keep the interactions. Calculating the interaction potential between pairs of amino acids is a useful tool [36, 37]. One can also build pairing preference matrices at protein-protein interfaces [38, 39]. Secondary structure data and information from correlated mutations in primary sequences have been used to predict protein residue contacts [29, 40]. Computational techniques, including neural networks, support vector regression and likelihood matrix have been used [38, 41, 42]. However some of the prediction methods have restrictions on input data, and the accuracy of all these prediction methods were still low. For example, the neural network

method can only achieve a prediction accuracy rate around 20%-30% and maximum likelihood around 20% for residue contacts. [29, 41].

Septin protein interactions and co-evolution

Co-evolution analysis is useful for those proteins that are somewhat conserved, but still show sequence variation. The septin protein sequences have sequence identities ranging from approximately 5% to 90%. As mentioned above, septins localize at many different parts of the cell, depending on the cell stage and organism. Septins can work as a scaffold to recruit other proteins. Thus, septins can interact with a variety of different proteins. Septins play important roles in cytokinesis and are considered to be a novel cytoskeletal component. The septum band in *Aspergillus nidulans* is a dynamic structure composed of actin, septin and formin [43]. Proteins involved with cytoskeleton proteins are good candidates for proteins that co-evolve with septins.

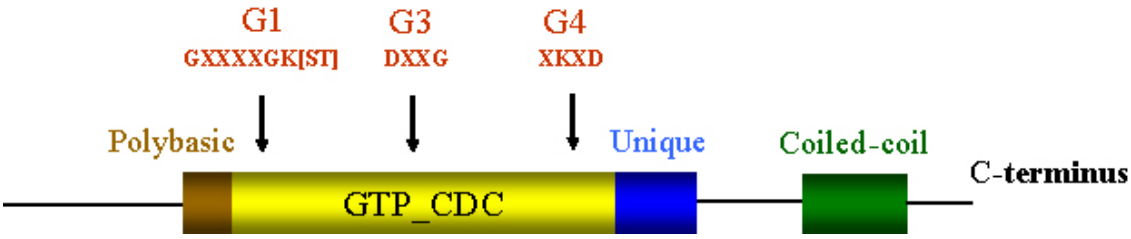
Experimental data are not available yet to show the exact interactions between septins and cytoskeleton proteins. Experimental techniques such as yeast two hybrid and immuno-colocalizations are used to find these interactions. Co-evolution analysis could computationally help to narrow down the possible interaction candidate proteins and more specifically the interacting sites, thus making it easier to design experiments.

Overview of this study

This dissertation includes three individual but related parts. Chapter 2 contains a phylogenetic analysis of the septin family. It suggests the classification of septins by sequence homology and identifies some conserved sites for septins. In Chapter 3, a new computational method designed to detect protein co-evolution and residue contacts is presented. It shows the use of mutual information for detecting correlated mutations in protein co-evolution and the design of a matrix

for predicting residues in close 3D distance. Chapter 4 presents work that is based on the previous two chapters, the applications of the techniques described in Chapter 3 to the septin family. We study co-evolution among different septin family members as well as between septins and proteins such as myosin and formin. Our studies suggested some interesting sites for further experimental analysis.

Figure 1.1. Septin sequence structure. The number of amino acids in septin sequences range from about three hundred to six hundred. Septins contain the conserved GTP_CDC binding domain with three motifs: G1, GxxxxGK[ST]; G3, DxxG; and G4, xKxD. Most septins have a short polybasic region preceding the G1 motif and a septin unique region after the G4 motif [4]. Some septins have a coiled-coil domain at the C terminus.


G1
GXXXXGK[ST]
G3
DXXG
G4
XKXD
Polybasic
Unique
Coiled-coil
C-terminus
GTP_CDC

CHAPTER 2

ANALYSIS OF SEPTINS ACROSS KINGDOMS REVEALS ORTHOLOGY AND NEW MOTIFS[1]

[1] Pan, F., Malmberg, R.L. and Momany, M. (2007) Analysis of septins across kingdoms reveals orthology and new motifs, *BMC evolutionary biology*, **7**, 103
Reprinted here with permission of publisher

ABSTRACT

Septins are cytoskeletal GTPase proteins first discovered in the fungus *Saccharomyces cerevisiae* where they organize the septum and link nuclear division with cell division. More recently septins have been found in animals where they are important in processes ranging from actin and microtubule organization to embryonic patterning and where defects in septins have been implicated in human disease. Previous studies suggested that many animal septins fell into independent evolutionary groups, confounding cross-kingdom comparison.

In the current work, we identified 162 septins from fungi, microsporidia and animals and analyzed their phylogenetic relationships. There was support for five groups of septins with orthology between kingdoms. Group 1 (which includes *S. cerevisiae* Cdc10p and human Sept9) and Group 2 (which includes *S. cerevisiae* Cdc3p and human Sept7) contain sequences from fungi and animals. Group 3 (which includes *S. cerevisiae* Cdc11p) and Group 4 (which includes *S. cerevisiae* Cdc12p) contain sequences from fungi and microsporidia. Group 5 (which includes *Aspergillus nidulans* AspE) contains sequences from filamentous fungi. We suggest a modified nomenclature based on these phylogenetic relationships. Comparative sequence alignments revealed septin derivatives of already known G1, G3 and G4 GTPase motifs, four new motifs from two to twelve amino acids long and six conserved single amino acid positions. One of these new motifs is septin-specific and several are group specific.

Our studies provide an evolutionary history for this important family of proteins and a framework and consistent nomenclature for comparison of septin orthologs across kingdoms.

BACKGROUND

Septins were first identified in the budding yeast *Saccharomyces cerevisiae* where they have been very well-characterized [2]. In *S. cerevisiae* five septins, Cdc3p, Cdc10p, Cdc11p, Cdc12p

and Shs1p, polymerize to form a ring at the mother-bud neck where they are important for bud site selection and cytokinesis. Two other yeast septins, Spr3p and Spr28p, are expressed during sporulation [44, 45]. Yeast septins have been shown to function as a scaffold organizing the division site and coordinating nuclear and cellular division. Septins have also been shown to act as a barrier, preventing diffusion of RNA and proteins between mother and daughter cells [2, 46].

Though not as well-characterized as those in yeast, septins in other fungi also appear to organize sites of cell division and new growth [47]. Septins have been found in a variety of animal tissues. In addition to acting as a diffusion barrier, animal septins are implicated in vesicle trafficking, apoptosis and cell movement [48]. In mammals septins appear to regulate membrane and cytoskeleton organization and abnormal septins have been linked with cancer and neurodegeneration [49-51].

Septins are P-loop GTPase proteins [19]. P-loop GTPases, including kinesin, myosin and ras proteins share at least five conserved motifs designated G1 to G5 within the GTP-binding domain [52]. The G1 motif, defined by the consensus element GxxxxGK[ST], forms a flexible loop which interacts with the phosphate group of the nucleotide [19, 52, 53]. The G2 motif is conserved within individual GTPase families, but not across the whole class [52]. The G3 motif contains several hydrophobic residues followed by DxxG [19, 52, 53]. This region binds Mg^{2+} and can interact with β and γ phosphates of GTP [19, 52, 54]. The G4 motif, NKxD, is important for GTP binding specificity [19, 55]. The G5 motif is found in some, but not all, members of the P-loop GTPase class [52].

Septins clearly contain the G1, G3 and G4 motifs [7] (Fig. 1). Septins purified from *Drosophila*, *Xenopus* and *Saccharomyces* have been shown to bind or hydrolyze GTP though the biological

significance of these activities and the specific functions of these motifs are not yet clear [56-58]. N-terminal to the GTPase domain, septins contain a polybasic region that has been shown to bind phosphoinositides [20, 59]. C-terminal to the GTPase domain, a 53 amino acid septin element conserved among many septins has been previously identified [60]. Most septins also contain a C-terminal extension predicted to form coiled-coils and shown to be needed for interactions between certain septins [20, 61, 62].

Previously fungal septins were placed into groups based on phylogenetic analysis [63] and mammalian septins were placed into groups based on primary sequence similarity [48]. Kinoshita [64] used phylogenetic analysis of two fungal yeast species and three animal species to conclude that orthologous relationships existed within fungal or animal septins, but not between fungal and animal septins making it impossible to compare model fungi and less tractable animals [64]. Recent genome projects provide an excellent opportunity to better understand the evolutionary relationships of septins. Here we identify 162 septins from 36 fungi, microsporidia and animals. Based on phylogenetic analysis we place the septins into five groups, two of which clearly contain orthologous fungal and animal septins. We also present three modified GTPase motifs, four new motifs and six individual amino acid positions which have been conserved among fungal, microsporidial and animal septins. Our results suggest that it should be possible to apply lessons learned from a subset of septins in model organisms to septins from mammals.

RESULTS

Database searches identified 166 septin-related sequences

We used the Cdc3p sequence of *Saccharomyces cerevisiae*, one of the best-studied septins, to query GenBank with the PSI-BLAST program and detected 876 sequences. From the PSI-

BLAST list we identified 166 unique potential septin sequences based on an e-value lower than e^{-3}, the presence of the G1, G3 and G4 GTPase motifs and other sequence similarities (Table 2.1). In our designation, the first three letters represent the species from which the sequence came (e.g. Sce represents *Saccharomyces cerevisiae*). Three septin sequences appeared to be truncated and were eliminated from further consideration (AbiSep, DyaSep2, and ZroCDC). We individually checked each of the remaining 163 potential septin sequences for the presence of the GTP_CDC domain [65].

Three of the 163 sequences were predicted to have only half of the GTP_CDC consensus domain and were designated "septin-like" (Gla, GzeHyp7 and NcrHyp7) (Table 2.1). In addition to septin sequences, our PSI-BLAST search with the Cdc3p query returned myosins and kinesins. A phylogenetic tree was built with representative septins, myosins, kinesins and ras GTPase family proteins to determine the relationship of the septin-like sequences to other GTPases. The three septin-like sequences did not group with any of the other GTPase superfamilies examined (data not shown). A BLAST search with the septin-like sequences did not give significant hits from any known protein families. This suggested that the septin-like sequences represent either ancient or diverged septins, or that they belong to an unknown protein family that shares some motifs with septins. The septin-like sequence found in *Giardia lamblia* is potentially illuminating for the evolution of this protein family because of *Giardia*'s position as a basal eukaryote.

The remaining 160 potential septin sequences grouped within a clade clearly separated from the other GTPase clades (data not shown). We designated these 160 sequences septins. After our PSI-BLAST search we became aware of two additional septins, human Sept 13 (HsaSept13) and *Ustilago maydis* Cdc10 (UmaCdc10), through reading of the literature [66, 67]. We included these sequences for a total of 162 septins. Consistent with previous reports, we found septins in

animals and fungi, but not in plants. Three septins were also found in the microsporidium *Encephalitozoon cuniculi.* We used a septin from *E. cuniculi* (EcuSepI, GenBank:gi|19075150) to query GenBank with PSI-BLAST a second time, and did not find any new potential septins.

Phylogenetic Analysis

Bayesian analysis of all septins: To investigate the evolutionary history of the septin gene family, we used the MrBayes program [68] to construct a phylogenetic tree for all 162 septins, rooting the tree with the *S. cerevisiae* myosin Myo2p. The septins could be grouped into five major clades (Fig. 2). Two clades contained fungal and animal septins (Groups 1 and 2); two clades contained fungal and microsporidial septins (Groups 3 and 4); one clade contained only fungal septins (Group 5). Group 1 consisted of two subgroups, 1A and 1B. Subgroup 1A further partitioned into one fungal clade and one animal clade supported by 0.96 credibility. The animal septins in Group 1A were closer to fungal Cdc10-type septins than to other animal septins. Group1A provides the strongest evidence for orthologous relationships between fungal and animal septins, suggesting the ancestral septin that gave rise to members of Group 1A originated before the fungal/animal split. Orthologous relationships between fungal septins in Group 2A and animal septins in Group 2B were supported with 0.78 credibility. Group 3 contained fungal and microsporidial septins. Though the credibility for Group 3 was only 0.55, all sequences except SpoSpn5, fell within a large clade with 0.85 credibility suggesting that the ancestral septin which gave rise to Group 3 arose before the fungal/ microsporidial split. Group 4 also contained fungal and microsporidial septins. Though it had a moderate credibility score of 0.76, sequences from Group 4 consistently fell within this clade. The small clade containing microsporidial septins EcuSep1 and EcuSep2 and fungal septin CalSpr3 had 0.98 credibility suggesting that the ancestral septin which gave rise to Group 4 also arose before the fungal/ microsporidial split.

Group 5, the smallest group, contained septins solely from filamentous fungi. The lack of orthologs from budding or fission yeast suggests that Group 5 septins either arose early in fungal evolution and were lost from yeasts or arose relatively late in fungal evolution.

Fungal septins: Ascomycetes with completed genome sequences had five to eight septins while basidiomycetes had four or five (Table 2.1). All fungi had single Group 1 and Group 2 septins. In contrast, at least some fungi had multiple Group 3, Group 4 and Group 5 septins. In particular, ascomycetous yeasts had three Group 3 septin paralogs. *U. maydis* and *Eremothecium gossypii* are the only two filamentous fungi in our study that lacked a Group 5 septin.

Animal septins: In the animals with completed genomes, nematodes had two or three septins, insects had four or five, fish had six and mammals had twelve or thirteen septins (Table 2.1). All animal septins fell within Groups 1 or 2, with Group 2B often showing the most expansion. The nematode *Caenorhabditis elegans* contained one septin from Group 1B and one from Group 2B. *C. briggsae* also had a single Group1B septin, but contained two Group 2B septins. The insect *Anopheles gambiae* contained a single Group 1B septin and three Group 2B septins, while *Drosophila melanogaster* had an additional Group 1B septin. The zebrafish *Danio rerio* contained a septin from Group 1A along with two septins from Group 1B and three from Group 2B. The mammals *Mus musculus* and *Rattus norvegicus* contained three Group 1A septins and five Group 2B septins. *Homo sapiens* contained three Group 1A septins and six Group 2B septins. *M. musculus* and *R. norvegicus* had five Group 1B septins while *H. sapiens* had four Group1B septins. Martinez and Ware previously divided mammalian septins into groups designated I-IV [48, 69, 70]; those groups fell within our Groups 1A, 1B and 2B as indicated in Figure 2.2.

Microsporidial septins: *E. cuniculi*, the single microsporidium with a completed genome included in our study, had three septins. *E. cuniculi* contained a single Group3 sequence and two Group 4 sequences. In contrast to all fungi and animals in our study, *E. cuniculi* contained no sequences from Groups 1 or 2.

Validation of tree topology using maximum likelihood: Maximum likelihood non-parametric bootstrapping is not ideal for large datasets; bootstrap values decrease as the taxon number increases [71] and the fast bootstrap methods without branch-swapping typically applied to large datasets may not be reliable at nodes with weak support [72]. None-the-less, because the nodes near the base of our Bayesian tree were weakly supported, we also constructed a phylogenetic tree using maximum likelihood methods. We used the PhyML program with 1024 bootstrap replicates to construct a phylogenetic tree with all 162 septins. The maximum likelihood tree gave the same basic tree topology as the Bayesian tree. For Groups 2, 3 and 5 maximum likelihood support values were similar to Bayesian support values (Fig.2, 78% versus 0.78, 51% versus 0.55 and 49% versus 0.55, respectively). For Group 4 the likelihood support value was higher than the Bayesian value (91% versus 0.76). For Group 1 the likelihood support value was much lower than the Bayesian support value (38% versus 0.8). However, support for Groups 1A and 1B was very similar by both methods (100% versus 0.96 and 100% versus 1.0).

Proposed names: Many of the septins we identified are listed as hypothetical proteins in GenBank or have been named after less related septins. We propose to name septins after the most closely related well-characterized septin within the same group (Table 2.1). The clades upon which proposed names are based are strongly supported and far from the base of the tree (Fig.2). Using this system, fungal and microsporidial septins from Groups 1-4 would be named Cdc3, Cdc10, Cdc11, Cdc12, Shs1, Spr3 or Spr28 after the most closely related *S. cerevisiae*

septin and those from Group 5 would be named AspE after the *A. nidulans* septin. The only exception would be fungal septins from *S. pombe*, which would continue to be called Spn1-7 because cell division cycle mutants bearing the Cdc name, but not correlating to the *S. cerevisiae* numbers, have been isolated independently. Mammalian and fish septins from Groups 1 and 2 would be named Sept1-13 after the human septins. Nematode septins would be named Unc59 and Unc61 after the *C. elegans* septins and insect septins would be named Pnut, Sep1, Sep2, Sep4, and Sep5 after *D. melanogaster* septins.

Domains and Motifs

To identify common motifs, we aligned all 162 septins and analyzed sequence patterns using the Weblogo program [24, 73]. In the following sections, septin amino acid positions are referenced to Cdc3p of *S. cerevisiae*.

GTPase domains: The G1 motif (GxxxxGK[ST]; SceCdc3 126-133) was the most conserved among the septin motifs (Table 2.2a). Glycines (G) were found in the first and sixth position in 99%-100% and in the fourth position in 94% of septins. Either K or R occupied the seventh position in 98%. All animal septins, all Group 1A fungal septins and all Group5 septins had a perfect consensus G1 motif. Eight fungal and microsporidial septins from Groups 2, 3 and 4 had derivatives of the consensus G1 motif (Supplementary Table 2.1). Our analysis also revealed that the two positions immediately following the G1 motif (SceCdc3 134-135) were [TS][LF] in 96%-97% of septins (Table 2.2a). A Prosite search using the extended G1 as query also identified other GTPases, so this extended G1 motif is not septin-specific.

The two consensus amino acids in the established GTPase G3 motif (DxxG; SceCdc3 204-207) were found in 83%-94% of septins. Our analysis also showed that the G3 motif consensus for septins could be further modified to DT[PV]GxG (SceCdc3 204-209) with each additional

position conserved in 86%-93% of septins (Table 2.2). Modified G3 motifs were found in all groups except for the animal and fungal Group 1A (Table S2.1).

In the G4 GTPase motif (NKxD, SceCdc3 286-289) N286 was often replaced by A, S, or G. K and D (SceCdc3 287 and 289) were found in 91% and 99% of septins, respectively. Perfect G4 consensus sequences were found in animal Group 1B and fungal Groups 2A and 4 and in fungi in Group 1A. Derived G4 motifs were found in fungal Groups 3 and 5 and in animal members of Group 1A and 2B. We also detected the pattern NxxPxI (SceCdc3 280-285) immediately upstream of the established G4 motif, with each of the three conserved amino acids in 91%-98% of septins. A Prosite search using this extended G4 pattern as query identified other GTPases, so it is not septin-specific.

Coiled-coil domains: The coiled-coil is a common structural motif that forms a super helix with heptad repeats and mediates protein-protein interactions [74, 75]. It exists in a broad range of proteins involved in numerous cellular processes [76]. Coiled-coil motifs have previously been identified at the C-terminus of the *S. cerevisiae* septins Cdc3 and Cdc12 where they are required for septin association and function [60]. A C-terminal coiled-coil has also been identified in Cdc11, but it is dispensable for function. Cdc10 is shorter than the other *S. cerevisiae* septins and lacks the C-terminal coiled-coil. We analyzed all 162 septin sequences for predicted coiled-coil domains using the COILS (http://www.ch.embnet.org/software/COILS_form.html) [77] and Multicoil programs (http://multicoil.lcs.mit.edu/cgi-bin/multicoil) [78]. Every member of the fungal Group 2A (Cdc3p) and the closely related animal Group 2B contained a predicted coiled-coil domain (Table 2.1). Similarly, all members of the fungal and microsporidial Group 4 (Cdc12p) contained the predicted coiled-coil. None of the animal or fungal septins in Group 1A (Cdc10p) had a predicted coiled-coil, while all the animal septins in the sister clade Group 1B

(M_II) had a predicted coiled-coil. None of the nine septins in the filamentous fungal Group 5 (AspE) were strongly predicted to have the coiled-coil; however, NcrHyp6, MgrHyp6, CneHyp5 and Gzehyp5 had weakly predicted coiled-coil domains (average probability across different window sizes <0.7, rather than 1). Though most members of the fungal and microsporidial Group 3 (Cdc11p) had a predicted C-terminal coiled-coil, five of the twenty-nine septins in Group 3 had no predicted coiled-coil (EcuSep3, CalSpr28, EgoHyp6, KlaHyp7 and SpoSpn7). Interestingly, the ascomycetes that have a Group 3 septin lacking a predicted coiled-coil contain two other Group 3 paralogs with predicted coiled-coils. However, the microsporidium *E. cuniculi* has only a single Group 3 septin.

New septin motifs: The Weblogo program assigned bitscores to amino acids in the established G1, G3 and G4 GTPase motifs ranging from a low of 2.7 (SceCdc3 position 204) to a high of 4.3 (SceCdc3 position 126). By considering relative frequency and using positions with bitscores above 2.7, we identified four new septin motifs and designated them Sep1- 4 (Table 2.2, b) and six new conserved single amino acid positions (Table 2.2, c). The Sep1 motif, ExxxxR (SceCdc3 position 237-242) is located between the established G3 and G4 domains (Figure 2.1) with each of the two consensus amino acids conserved in 96-98% of septins. A Prosite search of the NCBI protein database using the Sep1 motif returned many proteins that were neither septins nor GTPases. The Sep2 motif, DxR[VI]Hxxx[YF]F[IL]xP (SceCdc3 247-259) is located between the G3 and G4 GTPase domains (Figure 2.1). Each consensus amino acid was present in 88%-96% of septins. A Prosite search with the Sep2 motif identified only septins, but not all septins, making this motif potentially useful for identification of new septin sequences. Four septins with a P rather than a V or I at position 250 are all in the SceSpr28 subclade of Group3. The Sep3 motif, GxxLxxxD (SceCdc3 261-268), is between the G3 and G4 GTPase domains

(Figure 2.1). Each consensus amino acid was present in 86%-96% of septins. A Prosite search with the Sep3 motif returned GTPases including septins. In position 264, the hydrophobic L is often conservatively replaced by I or V. Only members of Group5 have the charged residue D at 264. The Sep4 motif, WG (SceCdc3 364-365), is in the C-terminus within the previously identified "septin unique element" and before the coiled-coil (Figure 2.1). The amino acids at these two positions were conserved in 92% of septins. A Prosite search with the Sep5 motif showed that it was also found in some other GTPases and hence is not septin-specific.

In addition to the four septin motifs, we detected six positions that contained single consensus amino acids in 86%-94% of septins (Table 2.2, c). One of these positions, upstream of the G1 GTPase motif in the polybasic region (SceCdc3 117; Figure 2.1), had a G in 99% of animal septins. In fungal septins it was moderately conserved except for four of the Group 5 septins where a P was substituted. The remaining five conserved single amino acid positions were after the G4 motif (SceCdc3 295, 300, 339, 360, and 396). In position 295, 94% of septins had the acidic residues D or E. However, in five septins from Group 5 the basic H residue was substituted.

Splice variants: Mammalian septins exhibit complex expression patterns and can produce a large number of splicing variants [67]. The human septin, SEPT9, spans a 240kb region, contains 17 exons, and is predicted to have 18 different transcripts encoding 15 polypeptides [79]. All of the conserved positions identified in our study were predicted to be retained in all variants encoded by SEPT9. Indeed, all splicing of human septin transcripts so far reported occurs in the regions encoding N- or C- termini and not in the regions encoding the conserved core of the protein.

DISCUSSION

Evolution

The origin of the septins in eukaryotes depends upon the interpretation of the septin-like sequence we found in *Giardia lamblia*. If this is considered a primitive septin, then a septin-like ancestor existed before the diplomonads arose. This septin-like ancestor was retained in the diplomonads, animals, fungi and microsporidia, but lost in plants. If the *G. lamblia* septin-like sequence is part of a separate GTPase family that shares some motifs with septins, then septins may have entered the common ancestor of animals, fungi and microsporidia via a horizontal gene transfer from bacteria, as proposed by Leipe [19].

Which ever origin is correct, our phylogenetic analysis suggests that septins might have evolved as follows (Figure 2.3): The ancestral septin sequence duplicated before the divergence of animals and fungi to become the ancestral Group 1 and Group 2 septins. The ancestral Group 1 septin duplicated and one paralog lost the C-terminal coiled-coil extension. Animals and fungi retained this shortened Group 1 paralog which gave rise to Group 1A septins. The longer paralog containing the C-terminal extension was lost from fungi, but retained in animals giving rise to Group 1B septins. Within fungal species there is a single Group 1 paralog, however in many animals, especially mammals, extensive duplication gave rise to multiple Group 1 paralogs. The ancestral Group 2 septin was retained in both animals and fungi giving rise to Group 2A and Group 2B septins. Fungi have single paralogs of Group 2 septins, while most animals, especially mammals, have multiple paralogs. In the lineage leading to fungi and microsporidia, the ancestral Group 1 and Group 2 septins duplicated giving rise to Group 3 and Group 4 septins. Unlike the single fungal paralogs of Group 1 and Group 2, Group 3 and Group 4 septins duplicated and diverged, giving rise to multiple paralogs, especially in the ascomycetes.

In the lineage leading to microsporidia, Group 1 and Group 2 septins were lost. This is consistent with recent views that microsporidia evolved from fungi [80]. Group 5 septins, found only in filamentous fungi, either arose early in fungal evolution and were lost in yeasts or arose relatively recently.

Motifs

Polybasic domain and Septin element: To be considered septin motifs, we required that sequences be at least as conserved as the GTPase motifs. While this stringent cut-off undoubtedly eliminated moderately-conserved or clade-specific sequences, it guaranteed the significance of identified positions. Only one amino acid (SceCdc3 117G) within the ten amino acid polybasic region previously shown to bind phosphoinositides (Figure 2.1; SecCdc3 110-120) [59] was conserved enough across all septins to be considered a septin motif in our analysis. Similarly, only 6 amino acids (sep1 motif and 2 conserved single amino acid positions) within the previously defined 53 amino acid "septin unique element" (SceCdc3 360-413) [62] meet our cut-off for septin motifs.

GTPase: Septins have been shown to bind and hydrolyze GTP [62]. Many lines of evidence suggest that guanine-nucleotide binding by septins is needed for their polymerization; however, low rates of nucleotide exchange and hydrolysis *in vitro* have led to questions about the significance of the GTPase activity. Consistent with the importance of guanine nucleotide binding for septin function, our analysis showed that the G1 GTPase motif, which forms the loop that interacts with the phosphate group of the nucleotide, and the G4 motif, which is important for GTP-binding specificity, were highly conserved, with 154 of 162 (95%) septins matching the respective consensus sequences (Supplementary Table 2.1). In contrast the G3 motif, which binds to the Mg2+ ion, matched the consensus for 135 of 162 (83%) septins.

Coiled-coil: In *S. cerevisiae* all septins except for Cdc10p (Group 1A) are predicted to have a C-terminal region containing a coiled-coil, a motif implicated in protein-protein interactions. Like Cdc10p, all Group 1A septins are missing the C-terminal region that contains the coiled-coil. Group 1B septins are all predicted to contain C-terminal coiled-coils. In elegant work, Versele and Thorner [60] showed that *S. cerevisiae* Cdc3p and Cdc12p associate through their C-termini and that Cdc11p and Cdc12p associate independently of their C-termini. In our analysis all Group 2 (Cdc3p) and Group 4 (Cdc12p) septins were predicted to contain C-terminal coiled-coils, while 5 of 29 Group 3 (Cdc11p) septins were not predicted to contain C-terminal coiled-coils. This pattern of conservation suggests that C-terminal coiled-coil interactions might be important for the association of all Group 2 (Cdc3p) septins with Group 4 (Cdc12p) septins while interactions outside the C-terminus might be important for the association of all Group 2 with Group 3 septins. Animals lack Group 4 septins, but Group 1B septins likely play the same role in polymerization by interacting with Group 2 septins. Indeed, mammalian Sept6 (Group 1B) and Sept7 (Group 2B) have been shown to interact via their C-termini leading Versele and Thorner to suggest that the Sept6-Sept7 complex is the animal counterpart of the Cdc3p-Cdc12p complex [62]. Group 5 septins, found in filamentous fungi, lack or have weakly predicted coiled-coils, suggesting that C terminal regions are not important for their interactions.

CONCLUSIONS

We analyzed 162 septins from microsporidia, fungi and animals. Septins were grouped into five classes, modified nomenclature based on these five classes was suggested and there was strong evidence for orthology between septins from different kingdoms. In addition to derivatives of already known G1, G3 and G4 GTPase motifs, four new motifs and six conserved single amino acid positions were identified. Though first discovered and best-studied in the yeast *S. cerevisiae*

it has become increasingly clear that the septins are important in animals. Earlier work based on septins from only five species suggested that there were no clear orthologs between the septins in fungal systems and those in mammals [64] confounding extrapolation from simple to more complex systems. With the availability of many more sequences, our work clarifies the relationships among septins and points to which comparisons are likely to be most informative.

METHODS

Database Searches

We used the 520-residue *Saccharomyces cerevisiae* septin protein Cdc3p (GenBank: gi|2507385) as the initial query sequence for PSI-BLAST searches against the non-redundant database (All non-redundant GenBank CDS translations+RefSeq Proteins+PDB+SwissProt+PIR+PRF) at NCBI [81] (http://www.ncbi.nlm.nih.gov). PSI-BLAST performs iterative profile searches by generating position specific scoring matrices to achieve high sensitivity. Three iterations were run with default parameters (Expect Value 10, Word Size 3, Blosum62, Gap Opening Penalty 11, Gap Extension Penalty 1, and With Inclusion Threshold 0.005) until no new septin or septin-like sequences were found. We examined each sequence retrieved from the PSI-BLAST output and removed duplicated and obviously incomplete sequences. We classified the remaining sequences as septins or septin-like proteins by examining the three GTP motifs of septins [64]: G1 (GxxxxGK[S/T]), G3 (DxxG) and G4 (xKxD) and their phylogenetic relationships with other septins.

Protein Alignments

We used CLUSTALX1.8 (ftp://ftp-igbmc.u-strasbg.fr/pub/ClustalX/) for protein multiple sequence alignment [82]. Default parameters were used, as no significant differences were observed when we tested different parameter combinations. Protein weight matrix Gonnet 250,

with Gap Opening Penalty 10 and Gap Extension Penalty 0.1 was used for pairwise alignments. Protein weight matrix Gonnet, with Gap Opening 10 and Gap Extension 0.2 was used for multiple alignments. We manually modified the multiple alignment output from ClustalX with the Bioedit program (http://www.mbio.ncsu.edu/BioEdit/bioedit.html). We used Weblogo Version 2.8.1 (http://weblogo.berkeley.edu/) to show the consensus structure of the sequences [24, 73]. Bit scores from the output were also used to help identify conserved regions.

Reconstruction of Phylogenetic Trees

We used MrBayes v3.1 (http://mrbayes.csit.fsu.edu/index.php) for phylogenetic analysis [68]. The amino acid model was estimated using the setting "aamodelpr=mixed" allowing the program to test and use the best fitting model for the dataset from 9 fixed rate protein models. We used 1,500,000 running generations, sample frequency of 200 and burn in period set to 40,000 to keep only the stationary phase samples. The chain number was set to 4 with 1 cold chain and 3 heated chains with heating coefficient λ=0.2. Two independent analyses were run simultaneously and converged. The consensus type was set to halfcompact. The myosin sequence from *Saccharomyces cerevisiae* Myo2p (gi|6324902) was used as outgroup. We also used PhyML [83] for maximum likelihood with bootstrap analysis of 1,024 replicates. The JTT amino acid substitution model was used. The proportion of invariant sites was estimated by maximizing the phylogeny likelihood. The number of relative substitution rate categories was set to 4 with gamma distribution parameter equal to 1. Tree topology, branch lengths and rate parameters were optimized.

Domain and Secondary Structure Predictions

We checked each sequence for domains with the Simple Modular Architecture Research Tool (http://smart.embl-heidelberg.de/) [65, 84]. An NCBI Conserved Domain Search was also used

[85]. Sequences were searched for coiled-coil domains with the COILS program at http://www.ch.embnet.org/software/COILS_form.html [77]; default parameters were used. Results from Multicoil were also considered (http://multicoil.lcs.mit.edu/cgi-bin/multicoil) [78]. Sequences with average probability above 0.7 were considered to have coiled-coil domains. Secondary structure was predicted using PSIPRED (http://bioinf.cs.ucl.ac.uk/psipred/psiform.html) [86].

AUTHORS' CONTRIBUTIONS

FP carried out the analysis and drafted the manuscript. RLM participated in the design of the study, helped in the analysis and helped to draft the manuscript. MM participated in the design of the study, helped in the data interpretation and helped to draft the manuscript and revise it critically.

ACKNOWLEDGEMENTS

This work was supported by NSF grant MCB 0211787 to MM and NIH grant 5R01GM072080-02 to RLM.

Figure 2.1. Typical septin structure. Septin sequences range from about three hundred to six hundred amino acids. Septins contain the conserved GTP_CDC binding domain with three motifs: G1, GxxxxGK[ST] (amino acids 126-135 in *S. cerevisiae* Cdc3p); G3, DxxG (amino acids 204-209 in *S. cerevisiae* Cdc3p); and G4, xKxD (amino acids 280-289 in *S. cerevisiae* Cdc3p). The previously described polybasic region (amino acids 110-120 in *S. cerevisiae* Cdc3p; [20, 60]) is shown as a black box and the previously described "septin unique element" (amino acids 360-413 in *S. cerevisiae* Cdc3p [60]); is shown as a grey box. S1-S4 mark positions of new septin motifs (Table 2.2b; amino acid 237-242, 247-259, 261-268, 364-365 in *S. cerevisiae* Cdc3p) and lines below diagram show conserved single amino acid positions (Table 2.2c; amino acids 117, 295, 300, 339, 360, 396 in *S. cerevisiae* Cdc3p). Many septins also have a predicted coiled-coil domain at the C-terminus (amino acids 476-507 in *S. cerevisiae* Cdc3p; [60]).

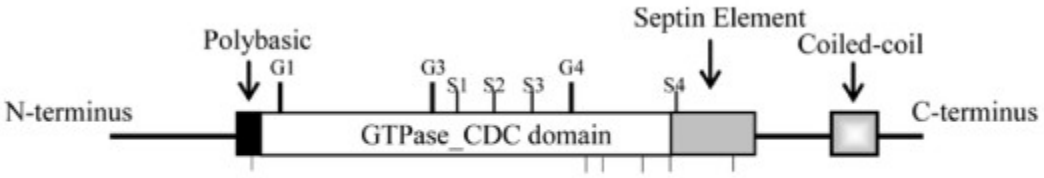

Septin Element
Polybasic
Coiled-coil
G1
G3
S1
S2
S3
G4
S4
N-terminus
GTPase_CDC domain
C-terminus

Figure 2.2. Overview phylogenetic tree of septin gene family. Half-compat consensus phylogram of 1.5 million generations of the MCMC analysis of the Bayesian phylogenetic analysis, discarding 400,000 generations as burn-in. Nodal numbers in front of the slash are posterior probabilities for Bayesian analysis. At the nodes where the tree topology agrees with the Bayesian analysis, numbers after the slash are bootstrap percentages from maximum likelihood bootstrap analysis using 1024 replicates. Red branches indicate animal lineage, green indicate fungal lineage and blue indicate microsporidia. Names in parenthesis under Group names indicate the best characterized fungal septin (CDC10, CDC3, CDC11 and CDC12, ASPE) or the mammalian septin classification of Martinez and Ware (2004) (MI, MII, MIII). See figures 2.3-2.5 for species names.

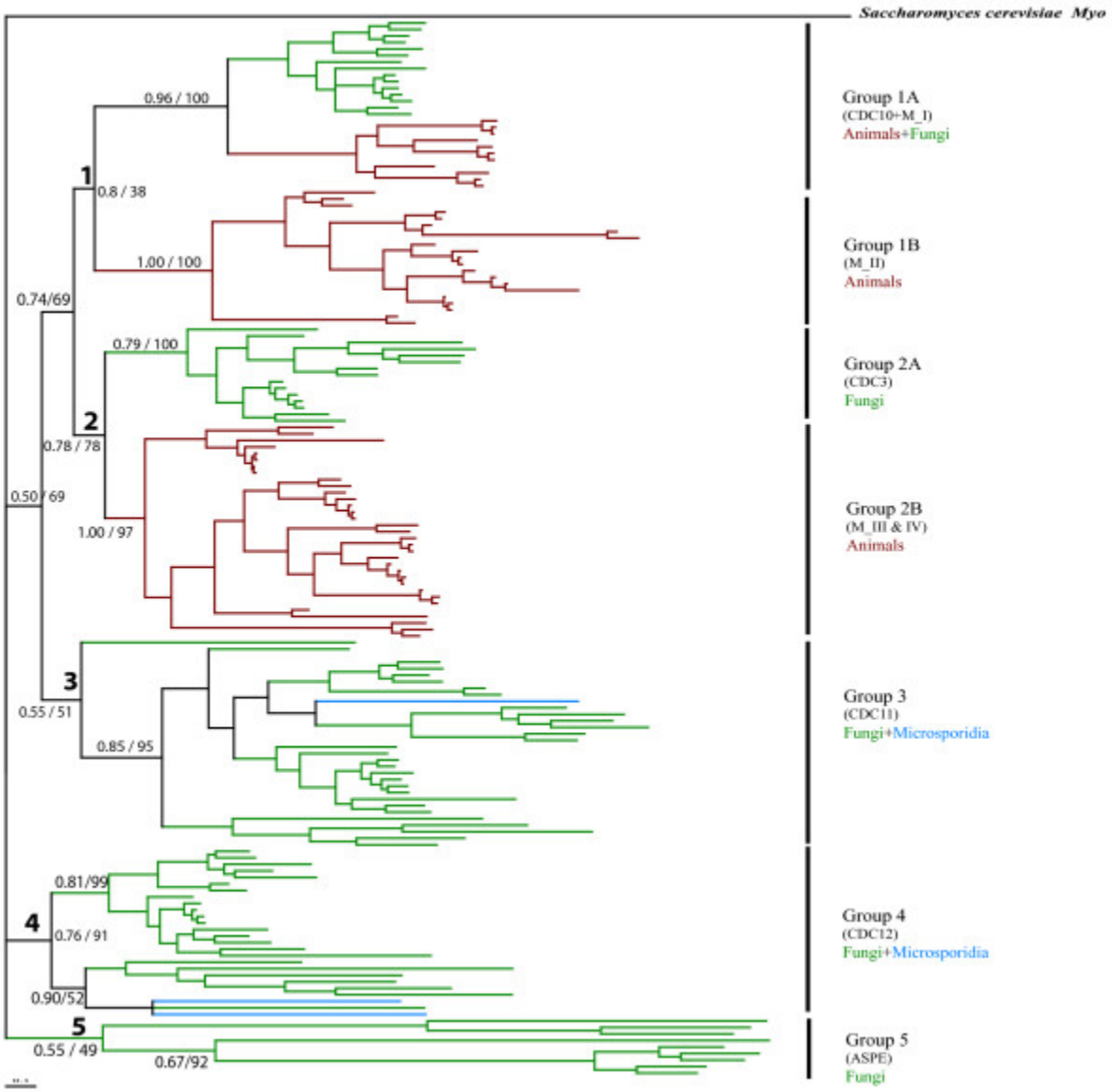

Saccharomyces cerevisiae Myo
0.96 / 100
1
0.8 / 38
1.00 / 100
0.74/69
0.79 / 100
2
0.78 / 78
0.50 / 69
1.00 / 97
3
0.55 / 51
0.85 / 95
0.81/99
4
0.76 / 91
0.90/52
5
0.55 / 49
0.67/92
Group 1A
(CDC10+M_I)
Animals+Fungi
Group 1B
(M_II)
Animals
Group 2A
(CDC3)
Fungi
Group 2B
(M_III & IV)
Animals
Group 3
(CDC11)
Fungi+Microsporidia
Group 4
(CDC12)
Fungi+Microsporidia
Group 5
(ASPE)
Fungi

Figure 2.3. Group 1 septin phylogenetic tree. Group 1 from figure 2.2 enlarged to show species names. Red branches indicate animal lineage and green indicate fungal lineage.

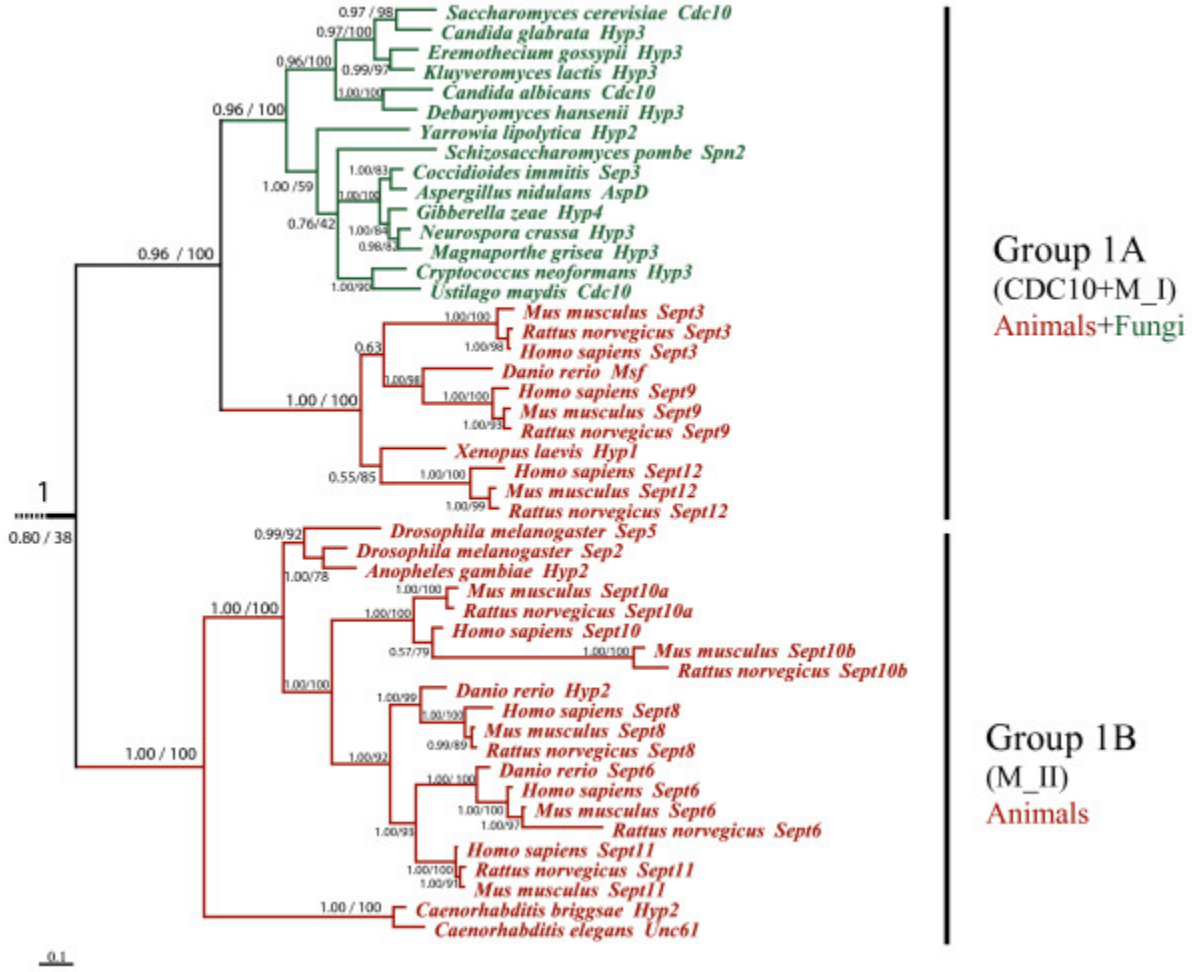

Saccharomyces cerevisiae Cdc10
Candida glabrata Hyp3
Eremothecium gossypii Hyp3
Kluyveromyces lactis Hyp3
Candida albicans Cdc10
Debaryomyces hansenii Hyp3
Yarrowia lipolytica Hyp2
Schizosaccharomyces pombe Spn2
Coccidioides immitis Sep3
Aspergillus nidulans AspD
Gibberella zeae Hyp4
Neurospora crassa Hyp3
Magnaporthe grisea Hyp3
Cryptococcus neoformans Hyp3
Ustilago maydis Cdc10
Mus musculus Sept3
Rattus norvegicus Sept3
Homo sapiens Sept3
Danio rerio Msf
Homo sapiens Sept9
Mus musculus Sept9
Rattus norvegicus Sept9
Xenopus laevis Hyp1
Homo sapiens Sept12
Mus musculus Sept12
Rattus norvegicus Sept12
Drosophila melanogaster Sep5
Drosophila melanogaster Sep2
Anopheles gambiae Hyp2
Mus musculus Sept10a
Rattus norvegicus Sept10a
Homo sapiens Sept10
Mus musculus Sept10b
Rattus norvegicus Sept10b
Danio rerio Hyp2
Homo sapiens Sept8
Mus musculus Sept8
Rattus norvegicus Sept8
Danio rerio Sept6
Homo sapiens Sept6
Mus musculus Sept6
Rattus norvegicus Sept6
Homo sapiens Sept11
Rattus norvegicus Sept11
Mus musculus Sept11
Caenorhabditis briggsae Hyp2
Caenorhabditis elegans Unc61
0.96 / 100
1.00 / 100
1.00 / 100
1.00 / 100
1
0.80 / 38
0.1
Group 1A
(CDC10+M_I)
Animals+Fungi
Group 1B
(M_II)
Animals

Figure 2.4. Group 2 septin phylogenetic tree. Group 2 from figure 2.2 enlarged to show species names. Red branches indicate animal lineage and green indicate fungal lineage.

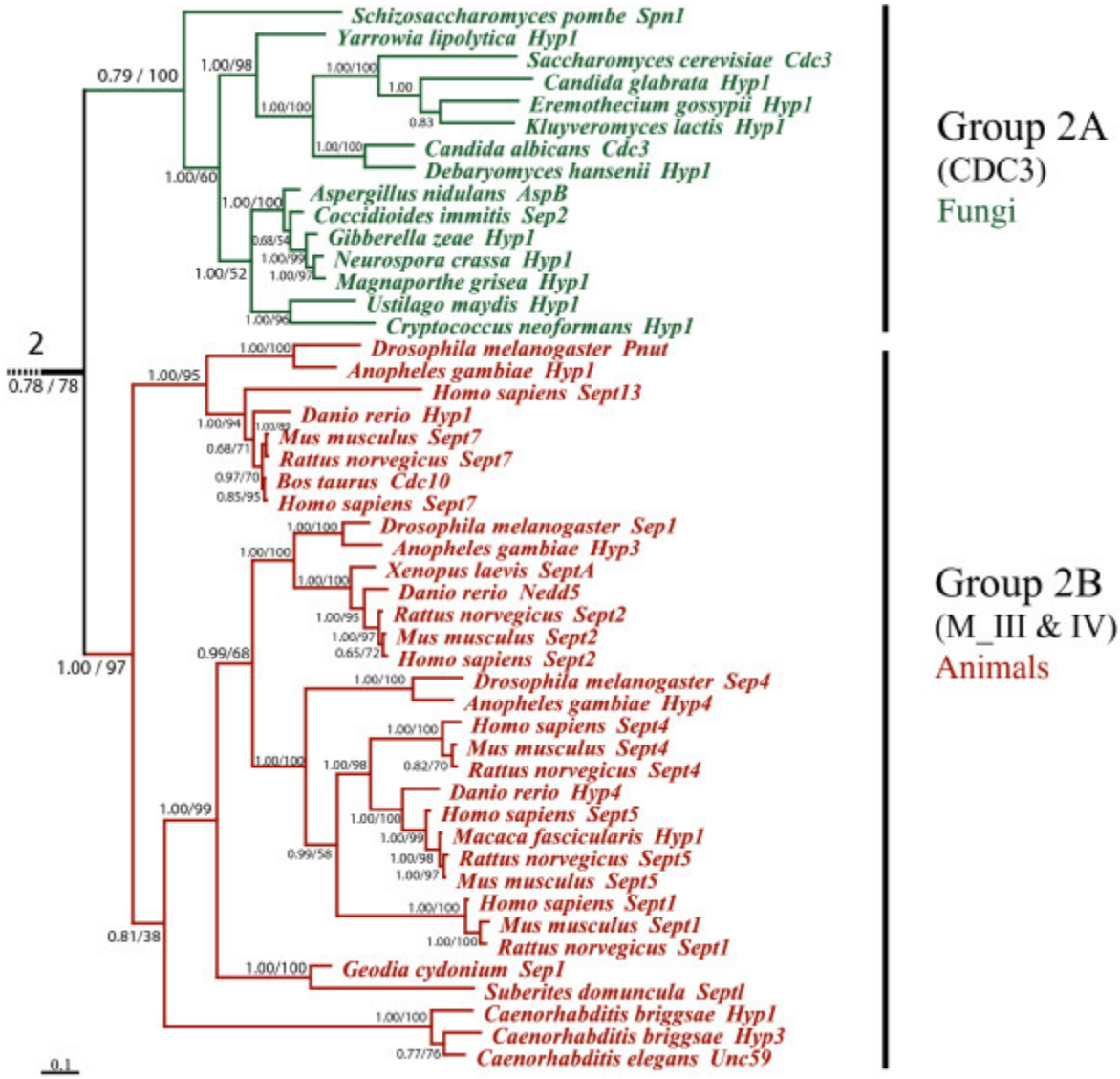

Schizosaccharomyces pombe Spn1
Yarrowia lipolytica Hyp1
Saccharomyces cerevisiae Cdc3
Candida glabrata Hyp1
Eremothecium gossypii Hyp1
Kluyveromyces lactis Hyp1
Candida albicans Cdc3
Debaryomyces hansenii Hyp1
Aspergillus nidulans AspB
Coccidioides immitis Sep2
Gibberella zeae Hyp1
Neurospora crassa Hyp1
Magnaporthe grisea Hyp1
Ustilago maydis Hyp1
Cryptococcus neoformans Hyp1
Drosophila melanogaster Pnut
Anopheles gambiae Hyp1
Homo sapiens Sept13
Danio rerio Hyp1
Mus musculus Sept7
Rattus norvegicus Sept7
Bos taurus Cdc10
Homo sapiens Sept7
Drosophila melanogaster Sep1
Anopheles gambiae Hyp3
Xenopus laevis SeptA
Danio rerio Nedd5
Rattus norvegicus Sept2
Mus musculus Sept2
Homo sapiens Sept2
Drosophila melanogaster Sep4
Anopheles gambiae Hyp4
Homo sapiens Sept4
Mus musculus Sept4
Rattus norvegicus Sept4
Danio rerio Hyp4
Homo sapiens Sept5
Macaca fascicularis Hyp1
Rattus norvegicus Sept5
Mus musculus Sept5
Homo sapiens Sept1
Mus musculus Sept1
Rattus norvegicus Sept1
Geodia cydonium Sep1
Suberites domuncula Septl
Caenorhabditis briggsae Hyp1
Caenorhabditis briggsae Hyp3
Caenorhabditis elegans Unc59
2
0.78 / 78
0.1
Group 2A
(CDC3)
Fungi
Group 2B
(M_III & IV)
Animals

Figure 2.5. Groups 3, 4 and 5 septin phylogenetic tree. Group 1 from figure 2.2 enlarged to show species names. Green branches indicate fungal lineage and blue indicate microsporidial lineage.

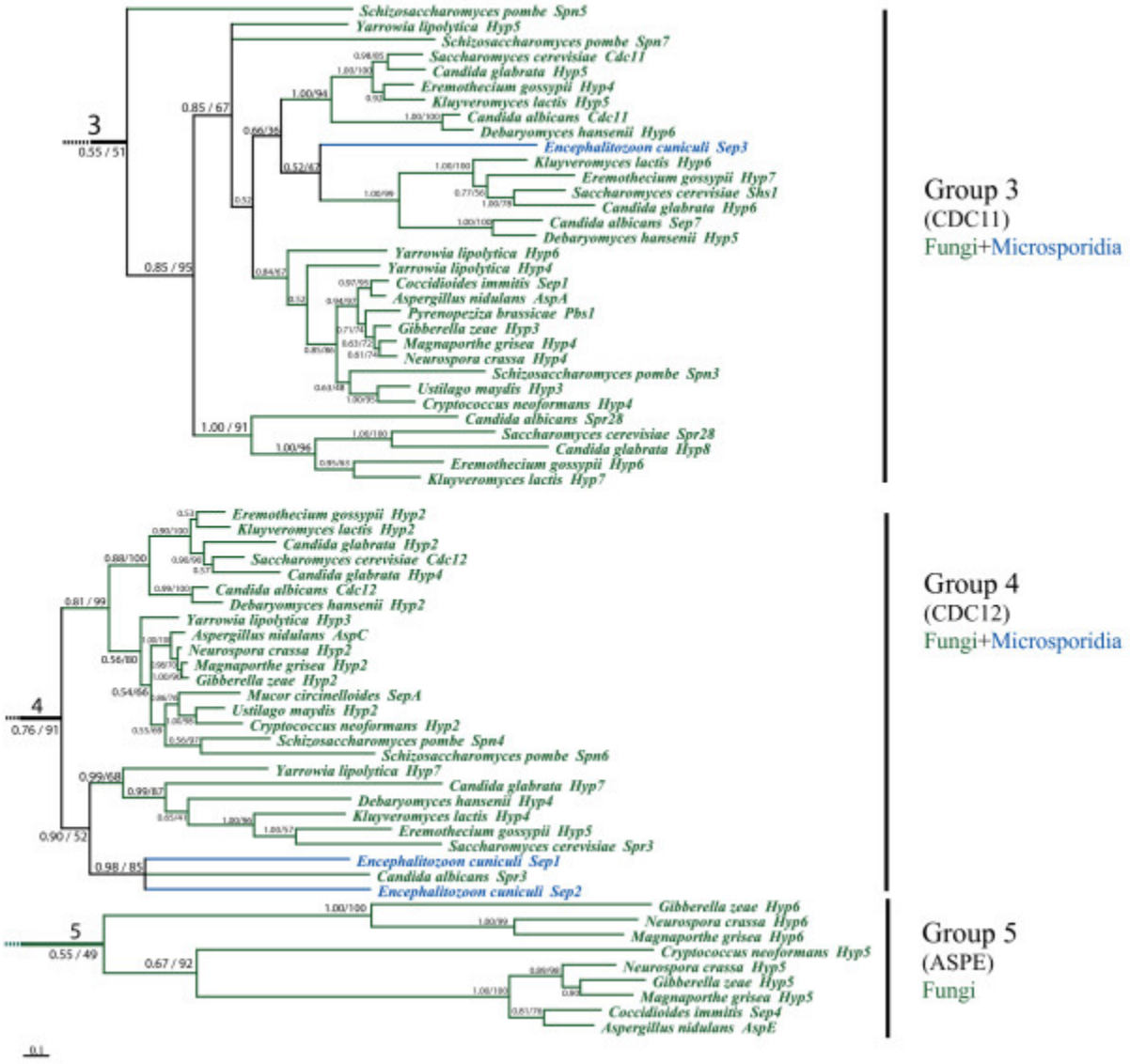

Schizosaccharomyces pombe Spn5
Yarrowia lipolytica Hyp5
Schizosaccharomyces pombe Spn7
Saccharomyces cerevisiae Cdc11
Candida glabrata Hyp5
Eremothecium gossypii Hyp4
Kluyveromyces lactis Hyp5
Candida albicans Cdc11
Debaryomyces hansenii Hyp6
Encephalitozoon cuniculi Sep3
Kluyveromyces lactis Hyp6
Eremothecium gossypii Hyp7
Saccharomyces cerevisiae Shs1
Candida glabrata Hyp6
Candida albicans Sep7
Debaryomyces hansenii Hyp5
Yarrowia lipolytica Hyp6
Yarrowia lipolytica Hyp4
Coccidioides immitis Sep1
Aspergillus nidulans AspA
Pyrenopeziza brassicae Pbs1
Gibberella zeae Hyp3
Magnaporthe grisea Hyp4
Neurospora crassa Hyp4
Schizosaccharomyces pombe Spn3
Ustilago maydis Hyp3
Cryptococcus neoformans Hyp4
Candida albicans Spr28
Saccharomyces cerevisiae Spr28
Candida glabrata Hyp8
Eremothecium gossypii Hyp6
Kluyveromyces lactis Hyp7
3
0.55 / 51
0.85 / 67
0.85 / 95
1.00 / 91
Group 3
(CDC11)
Fungi+Microsporidia
Eremothecium gossypii Hyp2
Kluyveromyces lactis Hyp2
Candida glabrata Hyp2
Saccharomyces cerevisiae Cdc12
Candida glabrata Hyp4
Candida albicans Cdc12
Debaryomyces hansenii Hyp2
Yarrowia lipolytica Hyp3
Aspergillus nidulans AspC
Neurospora crassa Hyp2
Magnaporthe grisea Hyp2
Gibberella zeae Hyp2
Mucor circinelloides SepA
Ustilago maydis Hyp2
Cryptococcus neoformans Hyp2
Schizosaccharomyces pombe Spn4
Schizosaccharomyces pombe Spn6
Yarrowia lipolytica Hyp7
Candida glabrata Hyp7
Debaryomyces hansenii Hyp4
Kluyveromyces lactis Hyp4
Eremothecium gossypii Hyp5
Saccharomyces cerevisiae Spr3
Encephalitozoon cuniculi Sep1
Candida albicans Spr3
Encephalitozoon cuniculi Sep2
4
0.76 / 91
0.81 / 99
0.90 / 52
0.98 / 85
Group 4
(CDC12)
Fungi+Microsporidia
Gibberella zeae Hyp6
Neurospora crassa Hyp6
Magnaporthe grisea Hyp6
Cryptococcus neoformans Hyp5
Neurospora crassa Hyp5
Gibberella zeae Hyp5
Magnaporthe grisea Hyp5
Coccidioides immitis Sep4
Aspergillus nidulans AspE
5
0.55 / 49
0.67 / 92
Group 5
(ASPE)
Fungi
0.1

Figure 2.6. Postulated septin evolution. Summary phylogeny of 31 species used in this study. This tree summarizes the evolution of the septins in the 31 organisms whose septins were identified and used in this study [50–53]. Red branches indicate animal lineages and green branches indicate fungal lineages. The table on the right of the tree indicates different groups of septin genes. Group 1 is red, group 2 is orange, group 3 is yellow, group 4 is green, and group 5 is blue. A triangle means the complete genome sequence was not available when the initial search was executed, so some septins might not have been identified due to incomplete sequence information.

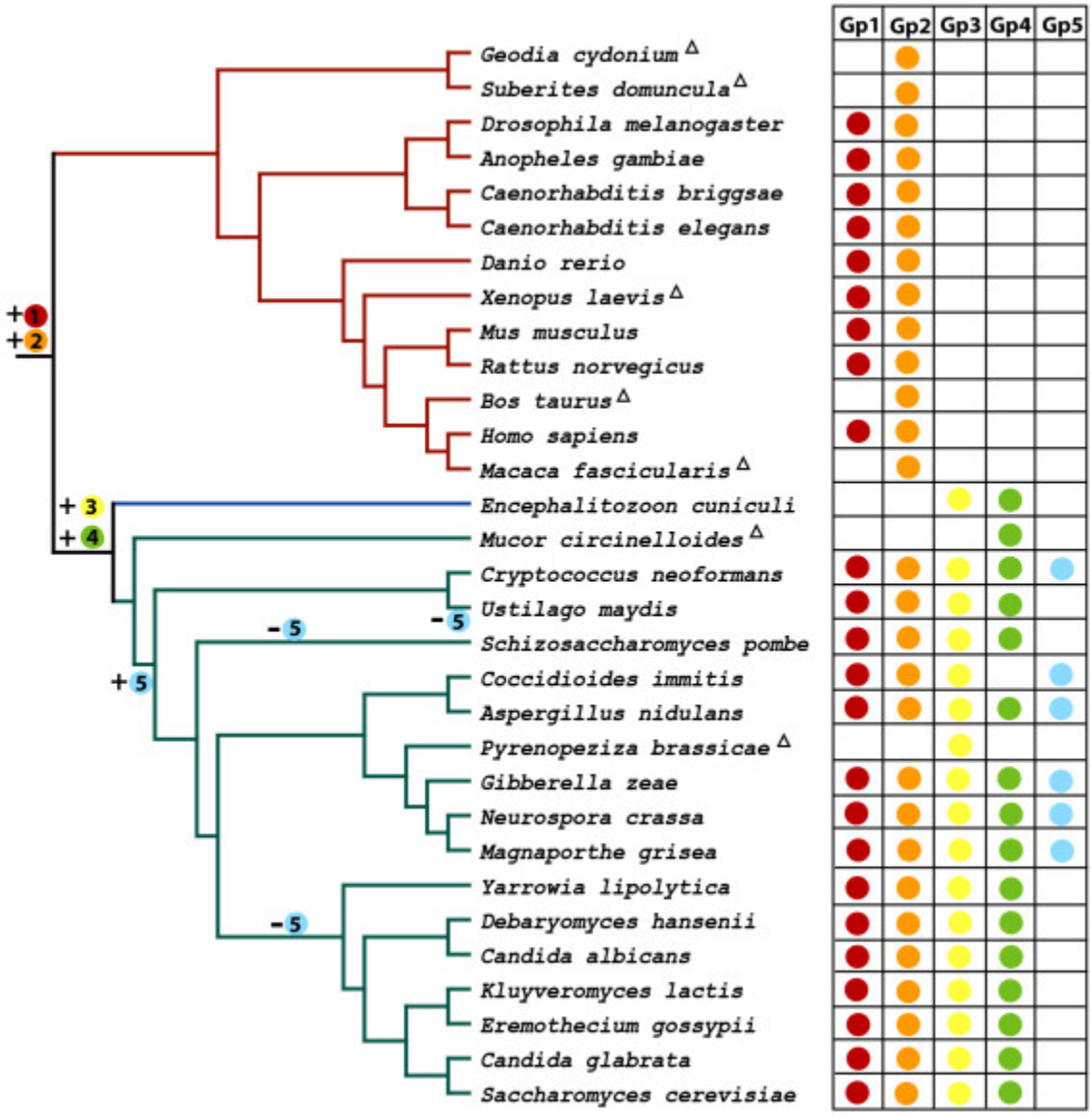

Gp1
Gp2
Gp3
Gp4
Gp5
Geodia cydonium Δ
Suberites domuncula Δ
Drosophila melanogaster
Anopheles gambiae
Caenorhabditis briggsae
Caenorhabditis elegans
Danio rerio
Xenopus laevis Δ
Mus musculus
Rattus norvegicus
Bos taurus Δ
Homo sapiens
Macaca fascicularis Δ
Encephalitozoon cuniculi
Mucor circinelloides Δ
Cryptococcus neoformans
Ustilago maydis
Schizosaccharomyces pombe
Coccidioides immitis
Aspergillus nidulans
Pyrenopeziza brassicae Δ
Gibberella zeae
Neurospora crassa
Magnaporthe grisea
Yarrowia lipolytica
Debaryomyces hansenii
Candida albicans
Kluyveromyces lactis
Eremothecium gossypii
Candida glabrata
Saccharomyces cerevisiae
+1
+2
+3
+4
+5
-5
-5
-5

Table 2.1. Septin and septin-like sequences analyzed.

GI[*]	Name in Paper[†]	Proposed Name[*]	Alias[‡]	Clade/ Species[§]		Group[¶]	G[‖]	C[**]
gi|2244629	AbiSep	/	sepA	F	Agaricus bisporus	Tr	+	+
gi|31198659	AgaHyp1	Pnut		A	Anopheles gambiae	2B	+	+
gi|31202059	AgaHyp2	Sep2		A	Anopheles gambiae	1B	+	+
gi|31206631	AgaHyp3	Sep1		A	Anopheles gambiae	2B	+	+
gi|31204715	AgaHyp4	Sep4		A	Anopheles gambiae	2B	+	+
gi|13398364	AniAspA	Cdc11		F	Aspergillus nidulans	3	+	+
gi|1791305	AniAspB	Cdc3		F	Aspergillus nidulans	2A	+	+
gi|34811845	AniAspC	Cdc12		F	Aspergillus nidulans	4	+	+
gi|34148975	AniAspD	Cdc10		F	Aspergillus nidulans	1A	+	-
gi|34811843	AniAspE	AspE		F	Aspergillus nidulans	5	+	-
gi|45645169	BtaCdc10	/		A	Bos taurus	2B	+	+
gi|39580970	CbrHyp1	Unc59b	CBG20268	A	Caenorhabditis briggsae	2B	+	+
gi|39589843	CbrHyp2	Unc61	CBG04550	A	Caenorhabditis briggsae	1B	+	+
gi|39584450	CbrHyp3	Unc59a	CBG19777	A	Caenorhabditis briggsae	2B	+	+
gi|17509405	CelUnc59	/		A	Caenorhabditis elegans	2B	+	+
gi|32566810	CelUnc61	/		A	Caenorhabditis elegans	1B	+	+
gi|729090	CalCdc3	/		F	Candida albicans	2A	+	+
gi|729064	CalCdc10	/		F	Candida albicans	1A	+	-
gi|21435770	CalCdc11	/		F	Candida albicans	3	+	+
gi|21435778	CalCdc12	/		F	Candida albicans	4	+	+
gi|21435802	CalSep7	Shs1	SHS1	F	Candida albicans	3	+	+
gi|46442449	CalSpr28	/		F	Candida albicans	3	+	-
gi|46444553	CalSpr3	/		F	Candida albicans	4	+	+
gi|50286825	CglHyp1	Cdc3		F	Candida glabrata	2A	+	+
gi|50284895	CglHyp2	Cdc12a		F	Candida glabrata	4	+	+
gi|50288449	CglHyp3	Cdc10		F	Candida glabrata	1A	+	-
gi|50289749	CglHyp4	Cdc12b		F	Candida glabrata	4	+	+
gi|50287113	CglHyp5	Cdc11		F	Candida glabrata	3	+	+
gi|50288341	CglHyp6	Shs1		F	Candida glabrata	3	+	+
gi|50291973	CglHyp7	Spr3		F	Candida glabrata	4	+	+
gi|50286937	CglHyp8	Spr28		F	Candida glabrata	3	+	+
gi|18476091	CimSep1	Cdc11		F	Coccidioides immitis	3	+	+
gi|24637104	CimSep2	Cdc3		F	Coccidioides immitis	2A	+	+
gi|24637108	CimSep3	Cdc10		F	Coccidioides immitis	1A	+	-
gi|24473756	CimSep4	AspE		F	Coccidioides immitis	5	+	-
gi|50257384	CneHyp1	Cdc3		F	Cryptococcus neoformans	2A	+	+
gi|50259101	CneHyp2	Cdc12		F	Cryptococcus neoformans	4	+	+
gi|50258769	CneHyp3	Cdc10		F	Cryptococcus neoformans	1A	+	-
gi|50257720	CneHyp4	Cdc11		F	Cryptococcus neoformans	3	+	+
gi|50260201	CneHyp5	AspE		F	Cryptococcus neoformans	5	+	-
gi|41055580	DreHyp1	Sept7	zgc:56383	A	Danio rerio	2B	+	+
gi|32822794	DreHyp2	Sept8	wu:fb22a03	A	Danio rerio	1B	+	+
gi|41152396	DreHyp4	Sept5	zgc:73218	A	Danio rerio	2B	+	+
gi|40538786	DreMsf	Sept9	MLL septin-like fusion	A	Danio rerio	1A	+	-
gi|45709377	DreNedd5	Sept2	zgc:63587	A	Danio rerio	2B	+	+
gi|47086783	DreSept6	/	zgc:66071	A	Danio rerio	1B	+	+
gi|50420949	DhaHyp1	Cdc3		F	Debaryomyces hansenii	2A	+	+
gi|50426961	DhaHyp2	Cdc12		F	Debaryomyces hansenii	4	+	+

gi|50425027	DhaHyp3	Cdc10		F	Debaryomyces hansenii	1A	+	-
gi|50426163	DhaHyp4	Spr3		F	Debaryomyces hansenii	4	+	+
gi|50418421	DhaHyp5	Cdc11		F	Debaryomyces hansenii	3	+	+
gi|50414330	DhaHyp6	Shs1		F	Debaryomyces hansenii	3	+	+
gi|730352	DmePnut	/		A	Drosophila melanogaster	2B	+	+
gi|17647925	DmeSep1	/		A	Drosophila melanogaster	2B	+	+
gi|17738071	DmeSep2	/		A	Drosophila melanogaster	1B	+	+
gi|24642597	DmeSep4	/	CG9699	A	Drosophila melanogaster	2B	+	+
gi|21356243	DmeSep5	/		A	Drosophila melanogaster	1B	+	+
gi|38047705	DyaSep2	/		A	Drosophila yakuba	Tr	-	+
gi|19075150	EcuSep1	Spr3a	ECU01_1370	M	Encephalitozoon cuniculi	4	+	+
gi|19074995	EcuSep2	Spr3b	ECU11_1950	M	Encephalitozoon cuniculi	4	+	+
gi|19173204	EcuSep3	Cdc11	ECU09_0820	M	Encephalitozoon cuniculi	3	+	-
gi|45198629	EgoHyp1	Cdc3		F	Eremothecium gossypii	2A	+	+
gi|45190841	EgoHyp2	Cdc12		F	Eremothecium gossypii	4	+	+
gi|45184824	EgoHyp3	Cdc10		F	Eremothecium gossypii	1A	+	-
gi|45191046	EgoHyp4	Cdc11		F	Eremothecium gossypii	3	+	+
gi|45199089	EgoHyp5	Spr3		F	Eremothecium gossypii	4	+	+
gi|45201271	EgoHyp6	Spr28		F	Eremothecium gossypii	3	+	-
gi|45185071	EgoHyp7	Shs1		F	Eremothecium gossypii	3	+	+
gi|14041182	GcySep1	/		A	Geodia cydonium	2B	+	+
gi|29249771	Gla	/		P	Giardia lamblia	Slk	-	-
gi|46121875	GzeHyp1	Cdc3		F	Gibberella zeae	2A	+	+
gi|46126005	GzeHyp2	Cdc12		F	Gibberella zeae	4	+	+
gi|46135811	GzeHyp3	Cdc11		F	Gibberella zeae	3	+	+
gi|46123315	GzeHyp4	Cdc10		F	Gibberella zeae	1A	+	-
gi|46128665	GzeHyp5	AspE		F	Gibberella zeae	5	+	-
gi|46122029	GzeHyp6	AspE2		F	Gibberella zeae	5	+	-
gi|46139179	GzeHyp7	/		F	Gibberella zeae	Slk	+	+
gi|16604248	HsaSept1	/		A	Homo sapiens	2B	+	+
gi|4758158	HsaSept2	/	Nedd5,Pnutl3,Diff6,KIA0158	A	Homo sapiens	2B	+	+
gi|22035572	HsaSept3	/		A	Homo sapiens	1A	+	-
gi|4758942	HsaSept4	/	H5,Bradion,Pnutl2,ARTS,MART,hCDCrel-2,Septin-M	A	Homo sapiens	2B	+	+
gi|9945439	HsaSept5	/	Pnutl,hCDCrel-1	A	Homo sapiens	2B	+	+
gi|22035577	HsaSept6	/	Sept2,KIA0128	A	Homo sapiens	1B	+	+
gi|4502695	HsaSept7	/	hCdc10	A	Homo sapiens	2B	+	+
gi|41147049	HsaSept8	/	KIA0202	A	Homo sapiens	1B	+	+
gi|6683817	HsaSept9	/	AF17q25, MSF, SepD1, Ov/Br septin, Pnutl4, KIA0991	A	Homo sapiens	1A	+	-
gi|18088518	HsaSept10	/	Sep1-like	A	Homo sapiens	1B	+	+
gi|8922712	HsaSept11	/	FLJ10849	A	Homo sapiens	1B	+	+
gi|23242699	HsaSept12	/	FLJ25410	A	Homo sapiens	1A	+	-
gi|113418512	HsaSept13	/		A	Homo sapiens	2B	+	+
gi|50306547	KlaHyp1	Cdc3		F	Kluyveromyces lactis	2A	+	+
gi|50309827	KlaHyp2	Cdc12		F	Kluyveromyces lactis	4	+	+
gi|50311269	KlaHyp3	Cdc10		F	Kluyveromyces lactis	1A	+	-
gi|50303889	KlaHyp4	Spr3		F	Kluyveromyces lactis	4	+	+

gi|50304439	KlaHyp5	Cdc11		F	Kluyveromyces lactis	3	+	+
gi|50311965	KlaHyp6	Shs1		F	Kluyveromyces lactis	3	+	+
gi|50311291	KlaHyp7	Spr28		F	Kluyveromyces lactis	3	+	-
gi|13358928	MfaHyp1	Sept5		A	Macaca fascicularis	2B	+	+
gi|38110101	MgrHyp1	Cdc3		F	Magnaporthe grisea	2A	+	+
gi|38106951	MgrHyp2	Cdc12		F	Magnaporthe grisea	4	+	+
gi|38109157	MgrHyp3	Cdc10		F	Magnaporthe grisea	1A	+	-
gi|38100755	MgrHyp4	Cdc11		F	Magnaporthe grisea	3	+	+
gi|38100224	MgrHyp5	AspE		F	Magnaporthe grisea	5	+	-
gi|38110686	MgrHyp6	AspE2		F	Magnaporthe grisea	5	+	-
gi|6453576	MciSepA	/	sepA	F	Mucor circinelloides	4	+	+
gi|8567344	MmuSept1	/	Diff6	A	Mus musculus	2B	+	+
gi|6754816	MmuSept2	/	Nedd5	A	Mus musculus	2B	+	+
gi|6755468	MmuSept3	/	mKIAA0991,G septin	A	Mus musculus	1A	+	-
gi|6755120	MmuSept4	/	M-Septin,H5	A	Mus musculus	2B	+	+
gi|6685763	MmuSept5	/	Cdcrel-1,Pnutl1	A	Mus musculus	2B	+	+
gi|20178348	MmuSept6	/		A	Mus musculus	1B	+	+
gi|9789726	MmuSept7	/	Cdc10	A	Mus musculus	2B	+	+
gi|39930477	MmuSept8	/	mKIAA0202	A	Mus musculus	1B	+	+
gi|28204888	MmuSept9	/	Sint1, E-septin, SLP-a	A	Mus musculus	1A	+	-
gi|26345492	MmuSept10a	/		A	Mus musculus	1B	+	+
gi|38082026	MmuSept10b	/	1700016K13Rik	A	Mus musculus	1B	+	+
gi|26324430	MmuSept11	/	D5Ertd606e	A	Mus musculus	1B	+	+
gi|20891621	MmuSept12	/	4933413B09Rik	A	Mus musculus	1A	+	-
gi|32417050	NcrHyp1	Cdc3		F	Neurospora crassa	2A	+	+
gi|32404966	NcrHyp2	Cdc12		F	Neurospora crassa	4	+	+
gi|32404320	NcrHyp3	Cdc10		F	Neurospora crassa	1A	+	-
gi|32422439	NcrHyp4	Cdc11		F	Neurospora crassa	3	+	+
gi|32417420	NcrHyp5	AspE		F	Neurospora crassa	5	+	-
gi|32411577	NcrHyp6	AspE2		F	Neurospora crassa	5	+	-
gi|32411845	NcrHyp7			F	Neurospora crassa	Slk	-	-
gi|5725417	PbrPbs1	Cdc11	pcd1	F	Pyrenopeziza brassicae	3	+	+
gi|34859284	RnoSept1	/	LOC293507	A	Rattus norvegicus	2B	+	+
gi|16924010	RnoSept2	/		A	Rattus norvegicus	2B	+	+
gi|9507085	RnoSept3	/	G-septin	A	Rattus norvegicus	1A	+	-
gi|32423788	RnoSept4	/	LOC287606,EG 3RVC,EG3-1RVC	A	Rattus norvegicus	2B	+	+
gi|16758814	RnoSept5	/	Gp1bb,CDCrel-1, PNUTL1ai, CDCrel-1AI	A	Rattus norvegicus	2B	+	+
gi|34932994	RnoSept6	/	LOC298316	A	Rattus norvegicus	1B	+	+
gi|12018296	RnoSept7	/	Cdc10	A	Rattus norvegicus	2B	+	+
gi|34870727	RnoSept8	/	LOC303135	A	Rattus norvegicus	1B	+	+
gi|13929200	RnoSept9	/	Slpa,E-Septin	A	Rattus norvegicus	1A	+	-
gi|34882181	RnoSept10a	/	LOC309891	A	Rattus norvegicus	1B	+	+
gi|34872099	RnoSept10b	/	LOC288622	A	Rattus norvegicus	1B	+	+
gi|34876531	RnoSept11	/	LOC305227	A	Rattus norvegicus	1B	+	+
gi|34868752	RnoSept12	/	LOC363542	A	Rattus norvegicus	1A	+	-
gi|6323346	SceCdc3	/		F	Saccharomyces cerevisiae	2A	+	+
gi|6319847	SceCdc10	/		F	Saccharomyces cerevisiae	1A	+	-
gi|6322536	SceCdc11	/		F	Saccharomyces cerevisiae	3	+	+

gi|6321899	SceCdc12	/		F	Saccharomyces cerevisiae	4	+	+
gi|6319976	SceShs1	/	Sep7	F	Saccharomyces cerevisiae	3	+	+
gi|6320424	SceSpr28	/		F	Saccharomyces cerevisiae	3	+	+
gi|6321496	SceSpr3	/		F	Saccharomyces cerevisiae	4	+	+
gi|19115666	SpoSpn1	/		F	Schizosaccharomyces pombe	2A	+	+
gi|19114071	SpoSpn2	/		F	Schizosaccharomyces pombe	1A	+	-
gi|13638491	SpoSpn3	/		F	Schizosaccharomyces pombe	3	+	+
gi|19114478	SpoSpn4	/		F	Schizosaccharomyces pombe	4	+	+
gi|19114952	SpoSpn5	/		F	Schizosaccharomyces pombe	3	+	+
gi|19075714	SpoSpn6	/	SPCC584.09	F	Schizosaccharomyces pombe	4	+	+
gi|15214304	SpoSpn7	/	SPBC19F8.01c	F	Schizosaccharomyces pombe	3	+	-
gi|20177379	SdoSeptl	/		A	Suberites domuncula	2B	+	+
gi|33302067	UmaCdc10	/		F	Ustilago maydis	1A	+	-
gi|46099680	UmaHyp1	Cdc3		F	Ustilago maydis	2A	+	+
gi|46099269	UmaHyp2	Cdc12		F	Ustilago maydis	4	+	+
gi|46099354	UmaHyp3	Cdc11	Sep3	F	Ustilago maydis	3	+	+
gi|34784614	XlaHyp1	Sept12	MGC68931	A	Xenopus laevis	1A	+	-
gi|12003372	XlaSeptA	Sept2		A	Xenopus laevis	2B	+	+
gi|50551445	YliHyp1	Cdc3		F	Yarrowia lipolytica	2A	+	+
gi|50549207	YliHyp2	Cdc10		F	Yarrowia lipolytica	1A	+	-
gi|50551749	YliHyp3	Cdc12		F	Yarrowia lipolytica	4	+	+
gi|50553330	YliHyp4	Cdc11a		F	Yarrowia lipolytica	3	+	+
gi|50549013	YliHyp5	Spr28		F	Yarrowia lipolytica	3	+	+
gi|50547965	YliHyp6	Cdc11b		F	Yarrowia lipolytica	3	+	+
gi|50557032	YliHyp7	Spr3		F	Yarrowia lipolytica	4	+	+
gi|13940377	ZroCDC	/	er001-c	F	Zygosaccharomyces rouxii	Tr	+	-

* Genbank identification numbers, http://www.ncbi.nlm.nih.gov .

† The first three letters represent genus and species names. The last letters represent current septin protein name.

* Proposed names based on first- or best-characterized septin in each clade.

‡Alias designations from Genbank and [66, 67]

§ A represents animals; F represents fungi; M represents microsporidia.

¶ Group names are assigned according to phylogenetic analysis shown in Figure 2.2. tr, truncated; slk, septin-like.

‖ Presence of full length GTP_CDC detected by the SMART program.

** Predicted coiled-coil at C terminus.

Table 2.2. Conserved motifs and single residues in septins.

a. Established Motifs and Extensions

Amino acid	Frequency (162 total)	Bitscore	Other Residues[+]
G1 motif (SceCdc3 126-135)			
G	162(100%)	4.3	/
x	/	/	/
x	/	/	/
G	152(94%)	3.6	N(4)
x	/	/	/
G	160(99%)	4.2	/
[KR]	158(98%)	3.4/ 0.4	/
[ST]	154(95%)	2/1.1	/
[TS]*	157(97%)	3.4/ 0.3	A(3)
[LF]*	156(96%)	1.9/1.2	M(3)
G3 motif (SceCdc3 204-209)			
D	135(83%)	2.7	E(6),N(4), S(4), L(3), T(3)
T*	141(87%)	3.1	A(7),S(5)
[PV]*	139(86%)	2.1/ 0.3	E(6), H(3)
G	153(94%)	3.8	N(4)
x*	/	/	/
G*	150(93%)	3.7	E(6)
G4 motif (SceCdc3 280-289)			
N*	156(96%)	3.8	T(5)
x*	/	/	/
x*	/	/	/
P*	159(98%)	4	L(2)
x*	/	/	/
I*	147(91%)	3.3	L(9), V(5)
x	/	/	/
K	148(91%)	3.4	R(11)
x	/	/	/
D	160(99%)	4	/

* Modifications of previously defined motifs.

[+] Most common examples of other residues; not all possibilities shown.

b. New Septin Motifs

Amino acid	Frequency	Bitscore	Other Residues[+]
Sep1 motif (SceCdc3 237-242)			
E	156(96%)	3.9	D(5)
x	/	/	/
x	/	/	/
x	/	/	/
x	/	/	/
R	158(98%)	3.9	T(2)
Sep2 motif (SceCdc3 247-259)			
D	156(96%)	3.9	E(2) ,G(2)
x	/	/	/
R	150(93%)	3.5	H(5)
[VI]	155(96%)	1.9/1	P(4)
H	153(94%)	3.6	D(4), Q(4)
x	/	/	/
x	/	/	/
x	/	/	/
[YF]	156(96%)	3.2/0.3	L(5)
F	147(91%)	3.3	L(9)
[IL]	153(94%)	2.9/0.3	V(8)
x	/	/	/
P	142(88%)	3.1	A(10), S(3)
Sep3 motif (SceCdc3 261-268)			
G	140(86%)	3	S(6)
x	/	/	/
x	/	/	/
L	151(93%)	3.2	D(5), I(4), V(3)
x	/	/	/
x	/	/	/
x	/	/	/
D	156(96%)	3.9	E(6)
Sep4 motif (SceCdc3 364-365)			
W	149(92%)	3.7	D(4)
G	149(92%)	3.6	/

[+] Most common examples of other residues; not all possibilities shown.

c. Single Conserved Position

SceCdc3 Position	Amino Acid	Frequency	Bitscore	Other Residues[+]
117	G	140(86%)	3.3	P(4)
295	[ED]	152(94%)	3.1/0.2	H(5)
300	K	150(93%)	3.6	R(9)
339	P	150(93%)	3.5	D(6)

360	R	150(93%)	3.4	/
396	T	150(93%)	3.4	S(5)

[+] Most common examples of other residues; not all possibilities shown.

Supplementary Table 2.1. Septin Derived GTPase Motifs*

a. G1 GTPase Domain: GxxxxG[KR][ST]

Groups	Sequence	Derived Motifs
2A	CalCdc3	GesglGKA
3	CalSpr28	GvndlGKK
2A	DhaHyp1	GesglGKK
4	EcuSep2	GrrglGTS
3	NcrHyp4	GasgtGES
3	PbrPbs1	GsslsPLV
3	SpoSpn7	GssytSYQ
3	YliHyp6	GpggsGRA

b. G3 GTPase Domain: DxxG

Groups	Sequences	Derived Motifs
3	CalCdc11	DtpN
2A	CalCdc3	TapG
3	CalSpr28	VtnN
4	CalSpr3	EtvN
4	CglHyp2	DtpA
3	CglHyp6	MtlG
3	CglHyp8	ImeG
2A	DhaHyp1	StpG
3	DhaHyp6	DtpN
2B	DreHyp1	(missing)
1B	DreSept6	NtvG
4	EcuSep2	TyhE
3	EgoHyp6	LapG
5	GzeHyp6	TrkR
1B	HsaSept10	NtvG
1B	HsaSept6	StvG
3	KlaHyp7	LipG
1B	MmuSept10a	NtvG
1B	MmuSept10b	KtvG
1B	MmuSept6	StvG
1B	RnoSept10a	NtvG
1B	RnoSept10b	KtvG
1B	RnoSept6	StvG
3	SceShs1	MthG
3	SceSpr28	LfpG
3	SpoSpn7	EvnG
3	YliHyp5	EgpG

c. G4 GTPase Domain: xKxD†

Groups	sequence	Derived Motifs
5	CneHyp5	sNvE
5	GzeHyp6	aRaD
1A	HsaSept12	aRaD
1A	Mmusept12	aRaD
5	NcrHyp6	sQaD
1A	RnoSept12	aRaD
2B	SdoSept1	iKcP
3	SpoSpn7	gNsN

*Upper case represents previously identified conserved positions.

† Loss of asparagine in the G4 motif NKxD is a feature for septin family, so the motif is represented by xKxD.

CHAPTER 3

PROTEIN CO-EVOLUTION AND ITS APPLICATIONS IN RESIDUE CONTACT PREDICTION[2]

[2] Fangfang Pan, Dongsheng Che, Michelle Momany, Liming Cai, and Russell L. Malmberg. Submitted to *Proteins:structure, function, bioinformatics*

ABSTRACT

Predicting protein 3D structure from primary sequence is a fundamental problem in computational biology, and building a residue physical contact map is a key step. The physical interactions between residues determine protein structure and functions, and those often involve highly conserved residues. However, residues can also mutate in a correlated way to preserve important interactions. Thus, co-evolving residues are one special group of residues worth studying for physical interactions. The rapid increase in the number of available protein sequences and structures allows the application of the mutual information statistic to protein co-evolution. We focused on 48 pairs of interacting proteins from the Protein Complex Crystallization Database and used mutual information to study residue co-evolution. Our analysis showed that, on average, co-evolving residue pairs were physically closer to each other than background. We derived protein co-evolving residue preference matrices using those co-evolving residues, and observed the existence of some individual residue pairing preferences. We also derived contact scoring matrices for co-evolving residues, and developed a new method to computationally predict whether co-evolving residue pairs are proximal or distant. Preliminary prediction results indicate the potential applicability of our approach to protein structure analysis.

INTRODUCTION

Predicting residue physical interactions in proteins with computational approaches has been of great interest as the experimental approaches, such as X-ray crystallography and NMR spectroscopy, are usually time consuming and have many limitations. It can help to reconstruct protein 3D structure if a residue physical interaction map is accurately predicted. Accurately identified residue physical contacts between proteins can predict important binding sites for protein-protein interactions.

Compared to the studies of general protein interactions, detecting exact interacting regions or even individual physically interacting positions is more challenging and usually requires more information. Previously secondary structures and the specific chemical and physical characteristics of the interface have been used to differentiate interacting interfaces from other protein surface regions[29, 40, 87]. Residue contact potentials (CP), derived from the known interactions, have also been used to indicate physical contacts between individual residues[39]. Computational techniques, such as neural network, support vector regression and likelihood matrices, were applied in the contact potential problem [38, 41, 42]. One promising approach is to consider the occurrence of co-evolution by looking at correlated mutations in proteins. Residues can co-evolve to maintain important interactions.

Correlations between pairs of residues have been the focus of RNA structure studies as functional RNA polymers evolved with the constraint of keeping stable internal base pairings rather than conserved primary sequences. Statistics from information theory were successfully applied in the co-variation study of ribosomal RNA [88]. Similar methods have also started to be applied exploring correlated mutations in homologous protein sequences [30, 34, 89-93]. Some

characteristics of co-evolving residues, including a smaller physical distance between them than average, have been observed [90]. Co-evolution has also been used to single out the right docking solution [34]. Thus, extracting correlated mutations by information theory and analyzing pairing preferences are potentially useful to predict contacts of co-evolving residues.

Recently, a large number of sequences have become available with the progress of genomics. This enables us to apply a mutual information (MI) based approach to study protein co-evolution, including co-variation between different proteins in a complex. In this paper, we introduce a new method to predict residue contacts from protein co-evolution. We built pairing preference matrices based on the co-evolving residues we identified by using information theory. We showed how to derive contact scoring matrices based on co-evolving pairing preferences. We propose a simple but novel prediction method using our contact scoring matrices. We believe that our method is complementary to other contact prediction approaches, as our method focuses mainly on co-evolving residues which are otherwise too variable to be discovered by current methods. We hope that our contact scoring matrices and prediction method can be used to detect some interacting residues of proteins.

METHODS

We summarize our method into two major stages. In stage 1, we collected the available data, aligned multiple sequences and analyzed mutual information (MI) (Figure. 3.1). In stage 2, we built co-evolving preference matrices, and proposed a new method for predicting proximal co-evolving residues. All procedures were automated except for manual adjustment of sequence alignments.

Collect protein-protein complexes

We focused our analysis on protein-protein complexes whose structures have been solved. The Protein Complex Crystallization Database (PCCD) contains 659 published unique protein–protein complexes [94]. Using the entries in PCCD, we retrieved sequences and structural information from the Protein Data Bank (PDB) (www.pdb.org). Many complexes have more than one identical chain for each polymer. Each chain was checked and only one chain from each set of identical chains in a polymer was kept to represent that polymer sequence.

Search for homologous sequences

For each protein complex, polymer sequences extracted from PDB by the above step were used as queries in BLASTP searches for homologs, run against the non-redundant database on NCBI [95]. Default parameters were used. There was usually more than one hit in an organism from one blast search due to the large size of some protein families. BLASTP results were sorted by organism names, and the sequence with the lowest *e*-value in each organism was used as the closest homolog of the query.

Filter protein complexes

We considered protein-protein complexes extracted above only when two requirements were satisfied: 1) The corresponding homologs from the same organism (iso-organismal homologs) existed for both chains; 2). The total number of organisms satisfying (1) was greater than 125. Figure 3.1 shows an example of extracting homologous sequences for a rat protein complex. By filtering 659 complexes in PCCD, we obtained 48 pairs of such complexes.

Align sequences and merge alignments of two polymers

ClustalW was used to align the homologous sequences of each chain [96]. The BioEdit program was used to aid the manual editing of alignments (http://www.mbio.ncsu.edu/BioEdit/bioedit.html). Protein alignments of any two chains from the same protein complex (protein with the same PDB ID) were merged to a single file so that MI could be calculated between and within chains at the same time (Figure 1).

Search for co-evolving residues

Mutual information has been applied in identifying co-evolving residues. It has been a concern in such analysis to separate phylogenetic influence from compensatory mutations [25, 36, 89, 93, 97]. To decrease the background noise, we used the approach of mutual information normalized by joint pair entropy to find co-evolving residues [90]. Briefly, for a given pair of columns (**x**, **y**) in a multiple alignment, we first calculated the values of entropy for column **x** and **y,** $H(\mathbf{x})$ and $H(\mathbf{y})$, the joint pair entropy $H(\mathbf{x}, \mathbf{y})$ and the mutual information $MI(\mathbf{x}, \mathbf{y})$ as follows:

$$H(\mathbf{x}) = -\sum_{i=1}^{20} p(x_i) \log p(x_i) \tag{1}$$

$$H(\mathbf{x},\mathbf{y}) = -\sum_{i=1}^{20}\sum_{j=1}^{20} p(x_i, y_j) \log p(x_i, y_j) \tag{2}$$

$$MI(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{20}\sum_{j=1}^{20} p(x_i, y_j)\log\frac{p(x_i, y_j)}{p(x_i)p(y_j)} \quad (3)$$

$$NMI = MI(\mathbf{x},\mathbf{y}) / H(\mathbf{x},\mathbf{y}) \quad (4)$$

where $p(x_i)$ and $p(y_j)$ are the probabilities of amino acid i and j respectively. We then considered the pair of column (**x**, **y**) was co-evolving only if $H(\mathbf{x})$ and $H(\mathbf{y})$ were greater than 0.3 and the *Z*-score (the units of standard deviation from the mean) of *NMI* was greater than 4 [90]. Using a criterion of Z score= 4 picks the top $3*10^{-5}$ positions with high mutual information, which means these positions have higher mutual information than the remaining 99.997% of positions in the dataset. We chose this *Z* score to be very stringent because of the large number of comparisons made.

For these predicted co-evolving residue pairs, we grouped them into two categories depending on whether two columns were from the same chain or two chains. For those co-evolving residues from the same chain, we called them type *W* (Within) co-evolving residues. Otherwise, we called them type *B* (Between) co-evolving residues. Figure 1 illustrates an example of two types of co-evolving residues.

Calculate physical distances of amino acids

The 3D coordinates of atoms in each polymer sequence were extracted from PDB to calculate the physical distance between any two amino acids. Given the 3D coordinates (*x*, *y*, *z*) of two atoms *A* and *B*, the physical distance between them was calculated as:

$$\vec{A} - \vec{B} = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2 + (z_A - z_B)^2} \quad (5)$$

The shortest distance between any pair of non-hydrogen atoms of two amino acids was used to represent the physical distance of these two amino acids.

Build residue pairing preference matrices

We obtained co-evolving positions as described in the previous section, and derived a 20 by 20 matrix to reflect the co-evolving pairing preferences of amino acids. We regarded pairing as symmetric; there was no difference between an (A, C) pair and a (C, A) pair, for example. Instead of using residue frequencies directly, we also took the residue compositions into consideration. We calculated the pairing likelihood between two amino acids *i* and *j* using the log odds:

$$A(i,j) = \log\left(\frac{p_{ij}^{Observed}}{p_{ij}^{Expected}}\right) = \log\left(\frac{p_{ij}^{o}}{p_{i}^{o} \times p_{j}^{o}}\right) = \log\left(\frac{O_{ij} \times N}{n_i \times n_j}\right) \quad (6)$$

where O_{ij} was the number of observed co-evolving residue pairs between amino acids *i* and *j* ($i, j = 1, 2, \ldots 20$), n_i (or n_j) was the number of co-evolving residue pairs with at least one residue being amino acid *i* (or *j*), and *N* was the total number of observed co-evolving residues. Depending on what type of co-evolving residues was used, we built three types of matrices by using co-evolving residues of type *W* only, type *B* only and all co-evolving residues.

Build contact scoring matrices

To determine which co-evolving residues were more likely to be physically close and possibly interact directly, we derived possible contact scoring matrices. Specifically, we first built two pairing preference matrices, one matrix built from co-evolving residues considered to be in contact (*i.e.*, the pair-wise distances of all co-evolving residues were within a physical distance cutoff *d*). The other matrix was built from those considered not in contact (*i.e.*, the pair-wise distances of all co-evolving residues were greater than *d*). We then calculated the score $s(i, j)$ in the contact scoring matrix as follows:

$$s(i, j) = A_{\leq d}(i, j) - A_{>d}(i, j) \quad (7)$$

Because there is no one single accepted distance cutoff to clearly define physical interactions of amino acids at the atomic level, we tested cutoff distances of 4Å, 5Å, 6Å, 7Å, 8Å to derive five contact scoring matrices.

Predict interacting co-evolving residues

After identifying co-evolving residues with the normalized mutual information approach, we predicted whether two positions in a multiple sequence alignment were in contact or not by evaluating the scores of co-evolving pairs from the contact scoring matrix. The average score (*AS*) was used to estimate whether two positions were in contact. It was defined as follows:

$$AS = \frac{1}{k} \sum_{i=1}^{k} s_i \quad (8)$$

where s_i was the score for a residue pair i, and k was the total number of sequences in the alignment. A suitable threshold T was chosen and compared with the value of AS to determine the co-evolving positions were in physical contact or not. The co-evolving positions were considered to be in contact if AS was less than T, otherwise, they were considered to be distant.

Validate prediction

We used a 10-fold cross validation approach to assess our contact scoring matrices and the prediction accuracy of our method. In particular, we divided all co-evolving residue data with distance information from 48 protein complexes into ten sets. Each time, nine sets were used as training data to build up a contact scoring matrix as described above. The remaining one set was used as testing data to see whether the co-evolving residues were in real contact or not. This process continued until all ten sets were evaluated. True positives (*TP*) were calculated as the number of co-evolving positions in contact predicted to be in contact; false negatives (*FN*) were

the number of co-evolving positions that are in contact but predicted not to be contact; true negatives (*TN*) were the number of distant co-evolving positions predicted to be distant; false positives (*FP*) were the number of distant co-evolving positions but predicted to be contact. Sensitivity (*Sens*) and specificity (*Spec*) were defined as follows:

$$Sens = \frac{TP}{TP + FN} \tag{9}$$

$$Spec = \frac{TN}{TN + FP} \tag{10}$$

RESULTS

Forty-eight qualified complexes

We checked the 659 protein-protein complexes from the Protein Complex Crystallization Database (PCCD), and narrowed our analyses to 48 protein-protein pairs after the filtering process as described in Methods (Figure 1). Each of these subunit pairs met our analysis requirements: each complex pair was different than the others and had a resolved protein 3D structure; the two polypeptide chains in one pair from the same complex were different; the number of iso-organismal homologs for both units in one complex pair was at least 125. Table 3.1 lists the PDB IDs and chain IDs of the 48 protein subunit pairs. A ribbon drawing example of complex 1AR1 is shown in supplementary Figure 3.1.

Co-evolving residues tend to be closer than average in physical distance

Co-evolving residues within a protein have been shown to often be close in 3D structure experimentally and in computational analysis [90, 98]. We tested whether this was true for our data including the co-evolving pairs that interact between proteins in one complex. Co-evolving residues within a protein sequence were denoted as type *W* (Within); while those two different

chains were denoted as type *B* (Between). We identified co-evolving residues from 48 multiple alignments as described in Methods. Supplementary Figure 3.2(a) shows that neighboring residues in the primary sequence were correlated and MIs of the neighboring residues were often high. We wanted to focus on the protein co-evolution constrained by protein 3D structure but not by the neighboring effects. Therefore, we excluded those co-evolving positions which were neighboring. In our study, we treated two columns were neighboring if they were separated by five residues or less.

We calculated the physical distances both of any two residue pairs (used as background) and of the co-evolving residue pairs from 48 protein-protein complexes. The physical distances between co-evolving residues could vary to a maximum 255 Å. The histogram in Supplementary Figure 3.3(a) shows the distance distribution of 7360 co-evolving residue pairs (including both type *W* and *B*). Among those data, 3152 co-evolving pairs were type *B* (Supplementary Figure 3.3(b)). While some co-evolving residues were physically too far away to have direct physical interactions, the average physical distance of co-evolving residue pairs was shorter than the average background distance. As shown in Table 3.2, the mean distance of all co-evolving residues was 33Å, compared with the distance of 40Å for background residues. The physical distance of all co-evolving residues was significantly shorter than that of the background using either *t*-test or Mann-Whitney-Wilcoxon test ($p<0.0001$). The physical distance of type *B* co-evolving residues was also significantly shorter when compared to the background distance calculated from all between-unit residues ($p<0.0001$). The mean distance between type *B* residues was 47Å compared to the background distance of 58Å (Table 3.3).

Though the mean distance was still too large to suggest physical interactions directly, our analysis indicated that the co-evolving residues were physically closer to each other than

random, including those residue pairs between two proteins. Thus, co-variation information of protein residues could possibly be used as a feature to identify the residues that are more likely to be in close contact, when combined with other information.

Amino acid compositions and pairing preferences of co-evolving residues

We first investigated whether there was any bias in amino acid composition for the co-evolving residues by calculating the composition percentage of each amino acid. Amino acid compositions were calculated for all 566,997 co-evolving residues in the multiple alignment files of 48 protein complexes. They were compared to the background composition of all 6,896,622 residues in the same alignment files. In general, there were no differences between the compositions of co-evolving residues and those of background residues (Figure 3.2). The correlation between the composition of co-evolving residues and the background was 0.95. This was similar to the correlations observed comparing residues at protein interfaces with the whole proteins, which were usually greater than 0.8 [39]. However, some amino acids, such as aspartic acid, cysteine, histidine, and asparagine, had slightly higher compositions in co-evolving residues. The two residues that had the greatest composition differences between co-evolving residues and the background were glycine and tyrosine, which had 15% less than and 47% more than the background composition (Supplementary Table 3.1).

A common method to represent pairing preferences is to build amino acid pairing matrices [38, 39, 99]. We used different types of co-evolving residues, *i.e.*, both type *W* and *B*, type *W* only and type *B* only to build 20 by 20 pairing matrices (Supplementary tables 3.2~3.4). These matrices show the preferences of amino acid pairs that are more or less likely to co-evolve together. We checked the co-evolving preferences in all three matrices. We found that main diagonal values were often positive for all co-evolving residues and type *B* co-evolving residues.

Among those, methionine, cysteine and lysine had very strong preferences to co-evolve with themselves (Supplementary Table 3.2). Off diagonal *B* type pair (F, W) and *W* type pair (Y, Q) also had very positive values. The preference differences from the type *B* only and type *W* only matrices were obvious, some extreme values could reach as low as -8 (Supplementary tables 3.3 and 3.4). Interestingly, most values of residue pairings between histidine and other amino acids were highly negative in type *B* matrix. Overall, there were some stronger preferences for particular co-evolving pairs.

Contact scoring matrices

Our analysis has shown that co-evolving residues could be physically distant as well as close, though the average distance of co-evolving residues was shorter than that of background. To detect the directly interacting residues, we divided co-evolving residues into two groups, a direct physical contact group and a more distant group. We then derived contact scoring matrices as described in Methods. Table 3.3 shows one scoring matrix using 8Å as a physical interaction cutoff. The score values in the matrix reflect the propensity of a pair of co-evolving amino acid residues to be physically close or distant. The higher the values in a table cell, the more likely those two co-evolving residues are in close contact. For example, the score of the C-W pair was 2.66, indicating they are highly likely found in physical contact. The score of the C-C pair was -7.18, indicating they are more likely found at a distance.

Cross validation of prediction

To evaluate the overall performance of our prediction method, we did a ten-fold cross validation for co-evolving residues in the multiple alignments for 48 protein complexes. We divided all co-evolving residue data into ten parts, and used nine parts to build a contact scoring matrix. The remaining one part was used to evaluate our prediction. We tested the data by using matrices

with different distance cutoffs (4Å, 5Å, 6Å, 7Å, 8Å), and evaluated the performance based on sensitivity and specificity.

We used different threshold values (T = -0.3, -0.2, -0.1, 0) to test our prediction performance. As we can see in Table 3.4, sensitivity decreased and specificity increased as the threshold value increased. For example, this method correctly identified 63% pairs with physical interactions and 32% of distant pairs with T = -0.3 and the 8Å matrix. In contrast, it could identify 25% contacting pairs and 71% non-contacting pairs with T = 0. Depending on different application goals, those parameters could be changed to best fit the problem.

DISCUSSION AND CONCLUSIONS

Correlated sequence changes are sufficient to detect the right interactions amongst many wrong alternatives within proteins [34]. We have presented a new approach to predict residue contact potentials based on information theory. Our analysis on co-evolving residues showed that the average 3D distance between co-evolving residues was shorter than average. Further analysis has shown that the existence of strong pairing preferences of residues, although there were few preferences in residue composition. We also built residue contact scoring matrices used to predict the closeness/distance for those co-evolving positions. Our idea of building scoring matrices for residue contact prediction is analogous to using BLOSUM or PAM matrices in BLAST searches for homologs. Validation of our predication method suggested the usefulness of our approach in predicting physically interacting residues.

Our derived contact scoring matrix and proposed prediction approach could be very helpful for predicting mutagenesis targets. Suppose two proteins are known to interact and one would like to know where the interaction residues/sites are. In addition to the conserved protein domains, the

researcher could first identify the co-evolving positions between these two proteins by using mutual information from multiple sequence alignment. The likelihood of two co-evolving positions to be in proximity could be predicted by applying our prediction scoring matrices. This could be used to help narrow down and differentiate the huge number of possible pairing residues for further mutagenesis analysis.

Our method is mainly based on mutual information in the multiple alignments. The use of mutual information requires a large number of sequences from different species in one alignment [90], leading to only a few number of qualified proteins. In our study, only 48 protein-protein complexes were qualified. In addition, the use of mutual information in multiple alignments for structure prediction relies on the assumption that the important protein 3D and quaternary structure be kept similar during evolution in different species, while the primary sequences show enough variation to be useful. This requires that in the protein alignment, the species are relatively close in their evolutionary relationship to keep the proteins physically of the same shape and functions.

Ideally, we need to use two types of co-evolving residues (*i.e.*, type *B* only and type *W* only) to derive two kinds of contact scoring matrices for predicting residue contacts. We analyzed the co-evolution both within and between different subunits in protein complexes, and found that there were some distinctive pairing preferences for these two groups. We thus believe that the type *B* contact matrix should be a better matrix used for predicting type *B* co-evolving positions, and similarly for the type *W* matrix. In our study, only 48 protein-protein complexes were used. There were not enough pairs of physically close co-evolving residues (such as 8Å) for type *B* or *W* alone to construct a contact scoring matrix. When there is sufficient data, a prediction matrix

for both types could be established and it might help to better predict the quaternary structures of some important proteins.

In summary, we have derived contact scoring matrices and developed a framework to predict physical contact of co-evolving amino acids. We hope that our matrix can more accurately capture the relationships of contact residues as more data become available, and our prediction can provide additional useful information to the traditional docking and threading methods.

GRANT SPONSOR

National Institutes of Health (5R01GM072080 to L.C. and R.L.M.); National Science Foundation (MCB 0211787 to M.M.)

AKNOWLEGEMENTS

The authors would like to thank Dr. Claiborne V. C. Glover III and Dr. Jaxk Reeves for helpful discussions.

Figure 3.1. An illustration of stage 1 of our method using a rat protein complex as an example. The rat protein complex has two interacting chains, 1 and 2. Only iso-organismal homologs were kept. Only five sequences in each alignment are shown here for simplicity. In the merged alignment, columns a and b are from the same chain, those co-evolving residues are denoted as type W (Within); while c and d are from two different chains, and thus they are denoted as type B (Between).

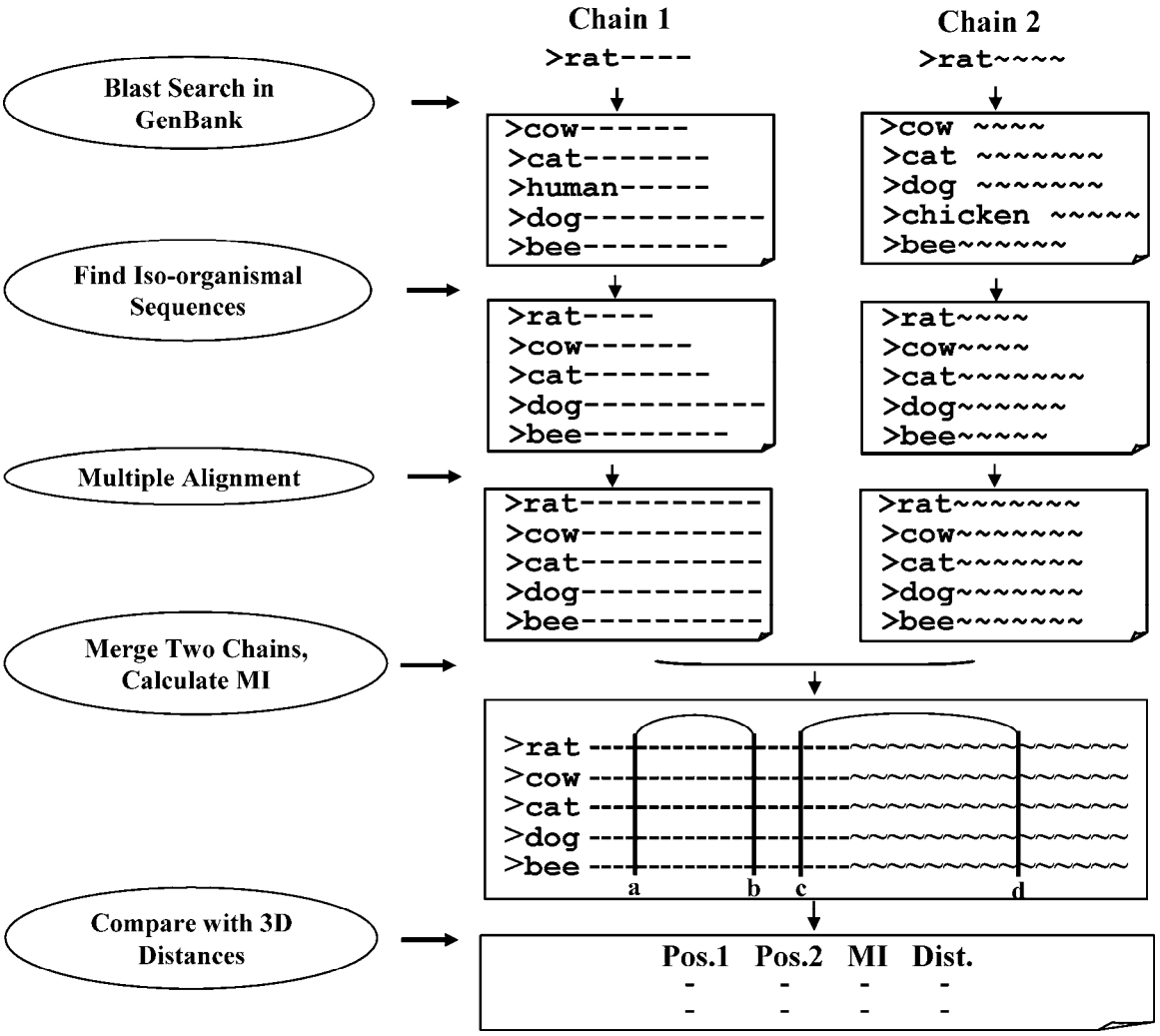

Chain 1
Chain 2
>rat----
>rat~~~~
Blast Search in GenBank
>cow------
>cat-------
>human-----
>dog----------
>bee--------
>cow ~~~~
>cat ~~~~~~~
>dog ~~~~~~~
>chicken ~~~~~
>bee~~~~~~
Find Iso-organismal Sequences
>rat----
>cow------
>cat-------
>dog----------
>bee--------
>rat~~~~
>cow~~~~
>cat~~~~~~~
>dog~~~~~~
>bee~~~~~
Multiple Alignment
>rat----------
>cow----------
>cat----------
>dog----------
>bee----------
>rat~~~~~~~
>cow~~~~~~~
>cat~~~~~~~
>dog~~~~~~~
>bee~~~~~~~
Merge Two Chains, Calculate MI
>rat
>cow
>cat
>dog
>bee
a
b
c
d
Compare with 3D Distances
Pos.1 Pos.2 MI Dist.

Figure 3.2. Amino acid compositions for all residues (background) and for co-evolving residues. The x-axis lists the twenty amino acid sorted by back-ground composition at increasing order. The y-axis is the composition percentage for each amino acid.

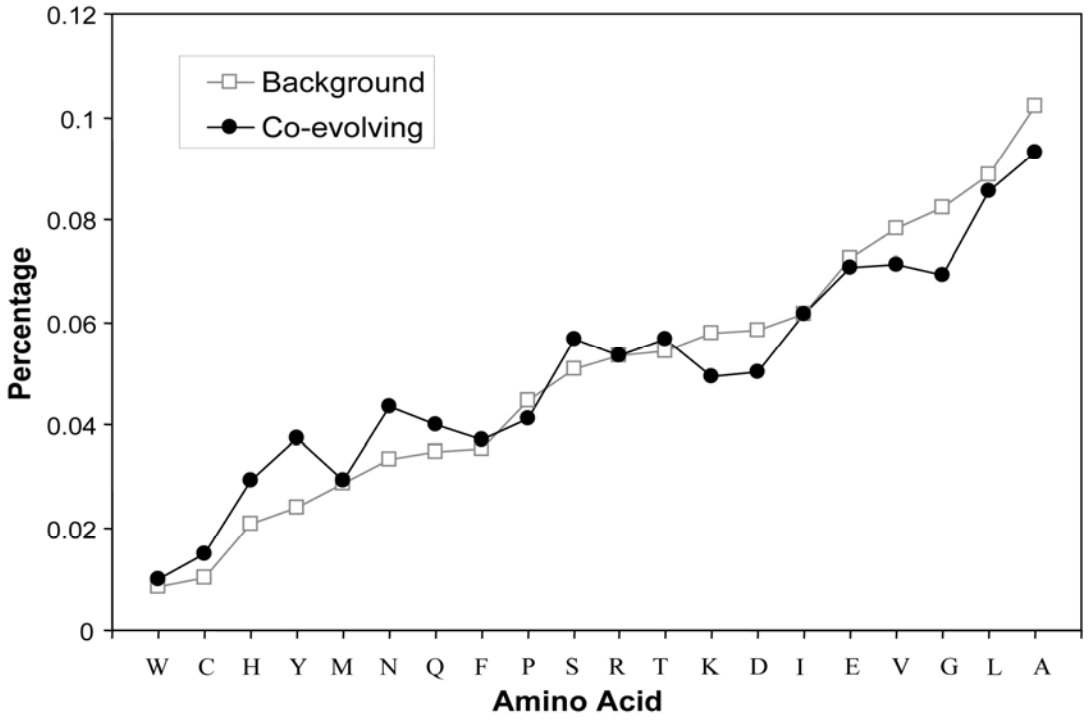

0.12
0.1
0.08
0.06
0.04
0.02
0
Percentage
Background
Co-evolving
W C H Y M N Q F P S R T K D I E V G L A
Amino Acid

Table 3.1. The PDB IDs and subunit designators of 48 hetero-complex pairs used in this study

1AIP_FH	1EFU_DC	1IXR_CB	1KKL_JC	1L0L_EA	1OKK_DA	1RJ9_BA	1W85_HG
1AR1_AB	1EZV_AD	1JW9_BD	1KYI_XR	1L0L_EB	1Q90_BC	1T9G_DS	1W85_JG
1DKG_DB	1EZV_AE	1JZD_BC	1L0L_BA	1MG9_AB	1QLA_ED	1TID_DC	1W85_JH
1E6E_DC	1F80_CF	1KB9_AD	1L0L_DA	1O94_BE	1QLE_AB	1TYG_GC	1W88_JG
1E7P_KJ	1GPW_EF	1KB9_AE	1L0L_DB	1O95_BF	1QLE_CA	1UBK_SL	1WDK_BD
1EBD_BC	1HT2_HD	1KF6_NM	1L0L_DE	1OFH_CN	1QLE_CB	1VF5_NQ	1XB2_AB

*The first four letters are the PDB ID and the last two letters are the 2 chain names in the complex as designated in PDB. Some protein complexes (eg. 1L0L) have multiple interacting chains and each interaction was considered as one pair

Table 3.2. Basic statistical results for physical distance analysis

	Number	Mean (Å)	Median (Å)	Maximum (Å)	P value
All co-evolving pairs	7360	33.31	29.65	254.87	
All background pairs	9133588	40.34	33.77	324.24	<0.0001
Type B co-evolving pairs	3125	46.83	43.28	254.87	
Background between pairs	3907261	57.63	48.18	324.24	<0.0001
Type W co-evolving pairs	4235	23.34	21.90	69.45	
Background within pairs	5226327	27.42	25.89	113.41	<0.0001

Type B: Between subunits in a protein complex

Type W: Within subunits in a protein complex

Table 3.3. Contact scoring matrix using 8 Å as a cutoff

-	I	V	L	F	C	M	A	G	T	S	W	Y	P	H	E	Q	D	N	K	R
I	-0.07	-0.08	0.19	-0.65	-0.39	-0.13	0.08	-0.34	0.01	-0.07	-0.64	-0.38	-0.13	-0.49	0.43	0.09	0.60	-0.10	-0.49	0.70
V	-0.08	-0.88	0.02	-0.27	-1.16	0.43	0.27	-0.04	0.65	0.38	-0.50	-0.23	-0.49	-1.18	-0.19	0.10	0.45	0.32	-0.36	-0.14
L	0.19	0.02	-0.97	0.32	-0.03	-0.21	-0.06	-0.13	0.98	-0.53	-5.07	-0.63	0.82	0.01	-0.52	-0.54	-0.11	0.49	0.33	0.14
F	-0.65	-0.27	0.32	-0.30	-0.84	-0.43	-0.28	-0.55	-0.62	-0.58	-1.58	-1.26	-0.47	0.39	0.64	0.22	0.62	-0.22	1.04	0.90
C	-0.39	-1.16	-0.03	-0.84	-7.18	-0.03	0.53	0.43	0.69	-0.48	2.66	-0.46	-5.20	0.58	-0.41	0.66	-0.50	-0.71	-0.67	-0.08
M	-0.13	0.43	-0.21	-0.43	-0.03	-0.08	-0.80	1.11	-0.11	-0.45	-0.90	0.38	0.05	-0.96	0.61	0.18	-1.71	-0.04	-1.14	-1.17
A	0.08	0.27	-0.06	-0.28	0.53	-0.80	-0.03	-0.32	0.37	-0.33	0.78	0.42	0.12	-0.46	-0.41	0.46	0.09	-0.04	-0.02	0.08
G	-0.34	-0.04	-0.13	-0.55	0.43	1.11	-0.32	-0.99	0.12	0.79	0.14	-0.06	0.43	-0.92	-0.25	0.48	-0.19	-0.02	0.37	-0.50
T	0.01	0.65	0.98	-0.62	0.69	-0.11	0.37	0.12	-1.54	-0.38	0.80	-0.08	-0.64	0.16	-1.39	0.23	0.23	0.60	-1.16	-1.27
S	-0.07	0.38	-0.53	-0.58	-0.48	-0.45	-0.33	0.79	-0.38	-0.97	0.03	0.35	0.04	-0.78	0.25	0.50	0.22	0.09	0.62	0.40
W	-0.64	-0.50	-5.07	-1.58	2.66	-0.90	0.78	0.14	0.80	0.03	-6.63	-0.93	-4.77	1.32	-0.52	-0.90	-0.90	-0.20	-0.92	-5.32
Y	-0.38	-0.23	-0.63	-1.26	-0.46	0.38	0.42	-0.06	-0.08	0.35	-0.93	-1.95	-0.50	-1.20	0.89	1.26	0.49	-0.83	0.53	0.06
P	-0.13	-0.49	0.82	-0.47	-5.20	0.05	0.12	0.43	-0.64	0.04	-4.77	-0.50	-1.63	-0.97	-0.21	0.41	0.28	0.71	-0.47	-0.70
H	-0.49	-1.18	0.01	0.39	0.58	-0.96	-0.46	-0.92	0.16	-0.78	1.32	-1.20	-0.97	-2.84	1.57	-1.72	-0.46	0.79	-0.73	1.50
E	0.43	-0.19	-0.52	0.64	-0.41	0.61	-0.41	-0.25	-1.39	0.25	-0.52	0.89	-0.21	1.57	-0.92	-0.06	-0.16	-0.88	0.69	0.03
Q	0.09	0.10	-0.54	0.22	0.66	0.18	0.46	0.48	0.23	0.50	-0.90	1.26	0.41	-1.72	-0.06	-1.71	-1.03	-0.55	-0.55	-1.51
D	0.60	0.45	-0.11	0.62	-0.50	-1.71	0.09	-0.19	0.23	0.22	-0.90	0.49	0.28	-0.46	-0.16	-1.03	-0.74	0.25	-0.04	0.32
N	-0.10	0.32	0.49	-0.22	-0.71	-0.04	-0.04	-0.02	0.60	0.09	-0.20	-0.83	0.71	0.79	-0.88	-0.55	0.25	-1.81	0.32	-0.02
K	-0.49	-0.36	0.33	1.04	-0.67	-1.14	-0.02	0.37	-1.16	0.62	-0.92	0.53	-0.47	-0.73	0.69	-0.55	-0.04	0.32	-0.71	-0.28
R	0.70	-0.14	0.14	0.90	-0.08	-1.17	0.08	-0.50	-1.27	0.40	-5.32	0.06	-0.70	1.50	0.03	-1.51	0.32	-0.02	-0.28	-1.23

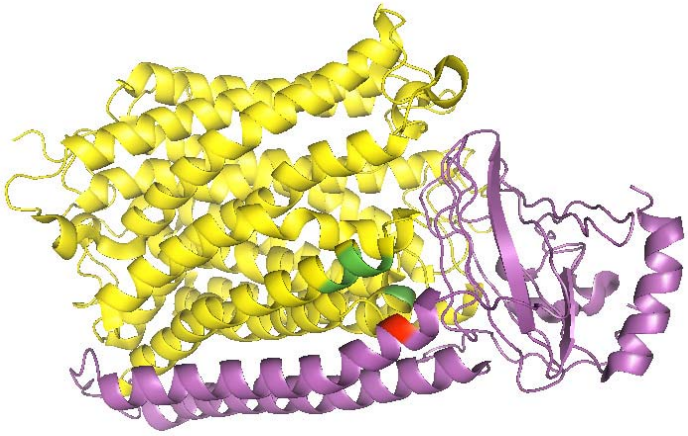
The residues are arranged according to their hydrophobicity, using the Kyte and Doolittle hydropathy index [100]

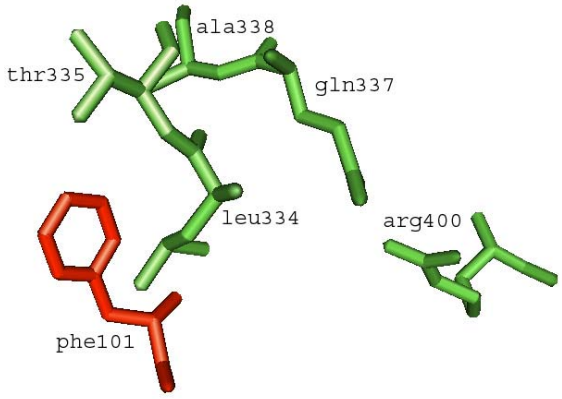
Table 3.4. Sensitivity and specificity for distance cutoff 4~8Å and score cutoff T -0.3~0

	***T*= -0.3**		***T*= -0.2**		***T*= -0.1**		***T*= 0**	
	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.
4Å	0.46	0.60	0.40	0.69	0.26	0.79	0.19	0.85
5Å	0.53	0.52	0.40	0.64	0.32	0.75	0.24	0.83
6Å	0.51	0.48	0.42	0.60	0.33	0.71	0.23	0.81
7Å	0.60	0.41	0.48	0.52	0.36	0.63	0.26	0.74
8Å	0.63	0.32	0.51	0.45	0.38	0.57	0.25	0.71

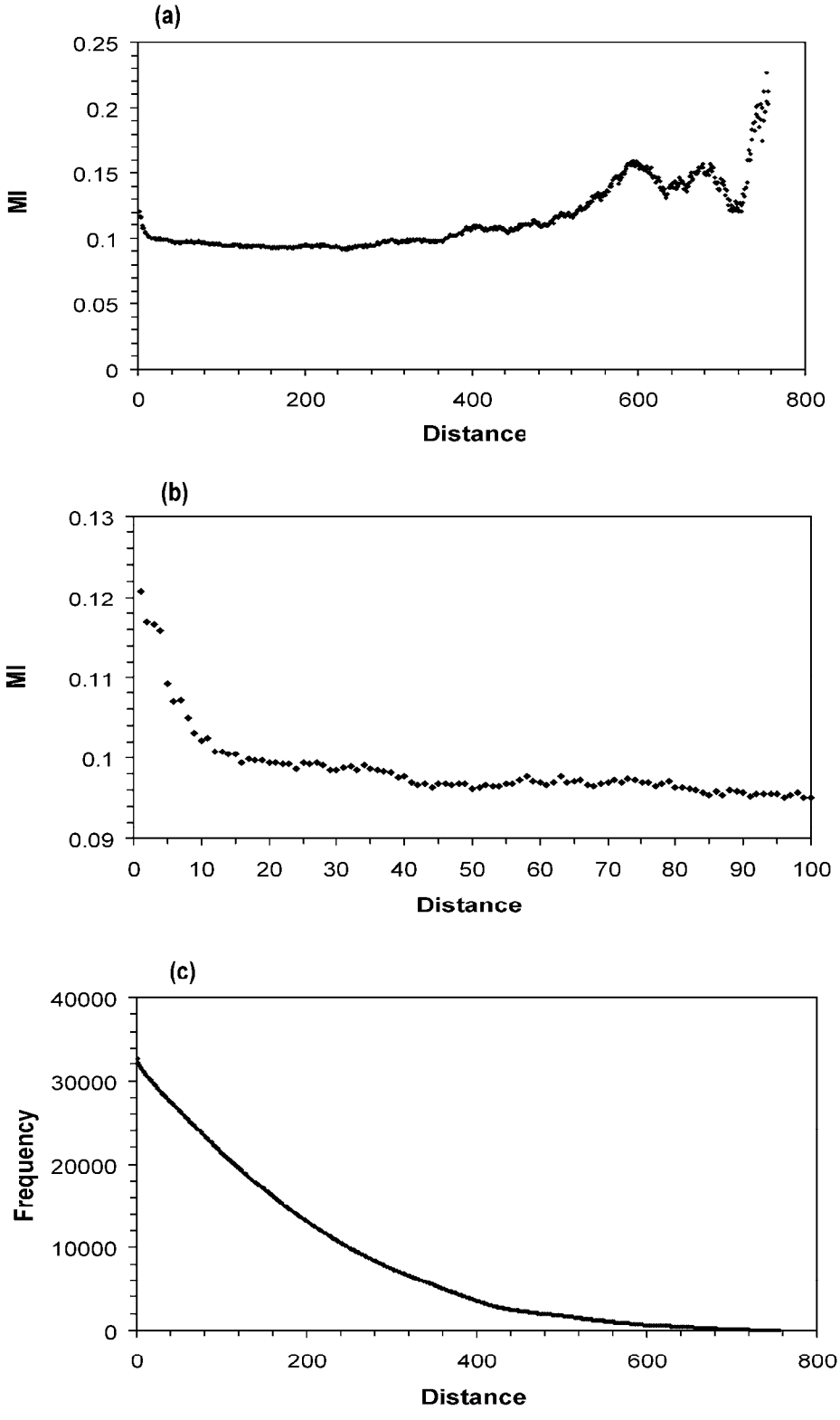
The median of all ten cross validation rounds was used in the corresponding cell to represent the overall sensitivity and specificity.

Supplementary Figure 3.1. A co-evolving and interacting network on cytochrome c oxidase. Graph was prepared with Pymol(DeLano, W.L. The PyMOL Molecular Graphics System (2002) DeLano Scientific, Palo Alto, CA, USA. http://www.pymol.org). We have found instances where high mutual information indicated co-evolution with multiple other residues in a constraint network. One example is cytochrome c oxidase (PDB_ID 1AR1). Cytochrome c oxidase (COX) is a large protein complex found in the mitochondrion and bacteria. It is an electron receptor at the terminal of electron transport chain in respiration [101]. We were able to identify 59 residue pairs that have high mutual information between units, out of which, 9 pairs had physical distance less than 8Å. One interesting subgroup is shown in Supplementary Figure 1. The Phe101(red) in subunit II has high mutual information with 5 different residues in subunit I, leu334, thr335, gln337, ala338 and arg400 (green). The site is far away from the Heme/Cu reaction center, but should be a very interesting site for further analysis.

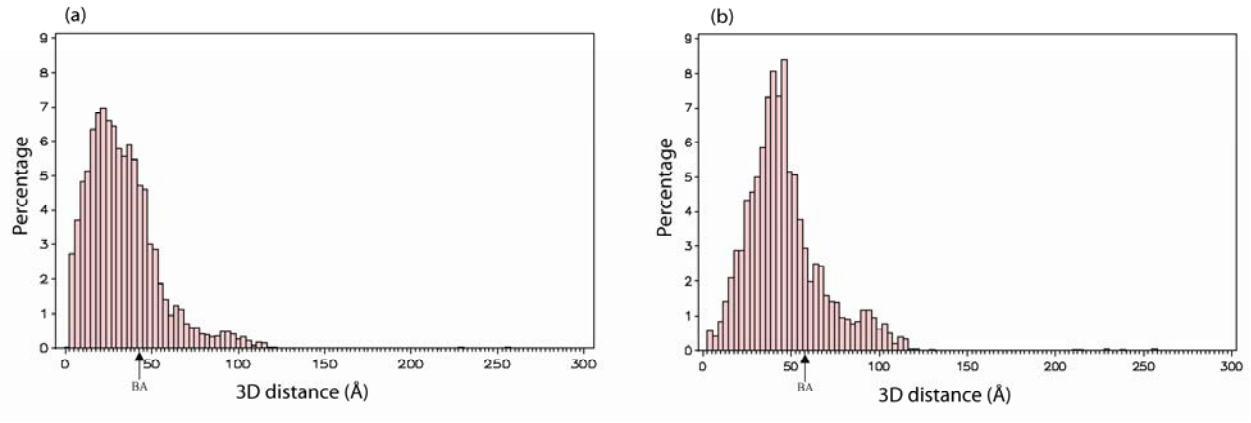



ala338
thr335
gln337
leu334
arg400
phe101

Supplementary Figure 3.2. MI distribution across residue distances in primary sequences. (a) The average normalized mutual information distribution. The x-axis is the linear primary sequence distance between any two residues within a protein sequence; the y-axis is the average normalized mutual information for all the residues of that distance category (eg. 1 residue apart, etc). (b) Enlarged part of figure 3(a) where the linear distance is from 1 to 100 residues apart. (c) The number of available MI data calculated for each distance category. The farther apart the two residues, the less data are available.


(a)
MI
0.25
0.2
0.15
0.1
0.05
0
0
200
400
600
800
Distance
(b)
MI
0.13
0.12
0.11
0.1
0.09
0
10
20
30
40
50
60
70
80
90
100
Distance
(c)
Frequency
40000
30000
20000
10000
0
0
200
400
600
800
Distance

Supplementary Figure 3.3. 3D distance histograms for residue pairs with high mutual information in 48 complexes. X-axis is the 3D distance between two residues with high mutual information, and y-axis is the percentage of residue pairs that fall in that distance category. Arrow points to the background average (BA) distance. (a) All residue pairs from both *B* and *W* types. (b) Residue pairs of *B* type.


(a)
Percentage
3D distance (Å)
BA
(b)
Percentage
3D distance (Å)
BA

Supplementary Table 3.1. Amino acid composition for co-evolving residues and background

	G	K	D	V	A	P	L	E	I	R	M	T	F	S	Q	W	N	H	C	Y
Background	0.082	0.058	0.058	0.078	0.102	0.045	0.089	0.072	0.062	0.053	0.029	0.054	0.035	0.051	0.035	0.008	0.033	0.021	0.010	0.024
Co-evolving	0.069	0.049	0.050	0.071	0.093	0.041	0.085	0.071	0.062	0.053	0.029	0.057	0.037	0.057	0.040	0.010	0.043	0.029	0.015	0.037
Difference	-0.013	-0.009	-0.008	-0.007	-0.009	-0.004	-0.003	-0.002	0.000	0.000	0.001	0.002	0.002	0.006	0.005	0.002	0.010	0.009	0.005	0.014
Difference Ratio	-0.159	-0.150	-0.140	-0.091	-0.088	-0.079	-0.037	-0.024	0.000	0.003	0.022	0.041	0.056	0.113	0.151	0.205	0.305	0.411	0.447	0.572

Supplementary Table 3.2. Symmetric amino acid pairing preference table for all co-evolving residues.

-	I	V	L	F	C	M	A	G	T	S	W	Y	P	H	E	Q	D	N	K	R
I	0.25	-0.11	0.04	0.06	0.03	-0.30	-0.30	-0.20	0.10	-0.24	-0.18	-0.27	-0.09	-0.25	-0.14	-0.16	-0.22	-0.08	-0.07	-0.32
V	-0.11	0.30	-0.06	-0.05	-0.02	-0.42	-0.06	-0.12	0.15	-0.22	-0.05	-0.55	-0.32	-0.06	-0.14	-0.32	-0.41	-0.23	-0.18	-0.17
L	0.04	-0.06	-0.18	-0.64	0.14	-0.30	0.03	-0.32	-0.14	-0.18	-0.57	-0.20	0.12	-0.38	0.04	0.09	0.08	-0.15	0.02	-0.45
F	0.06	-0.05	-0.64	0.52	-0.02	-0.60	-0.04	0.21	-0.08	-0.33	0.78	-0.30	-0.29	-0.01	0.08	0.08	-0.05	-0.34	-0.23	-0.01
C	0.03	-0.02	0.14	-0.02	1.91	-0.28	0.03	0.14	-0.57	-0.36	0.17	-0.34	-0.59	0.23	-0.20	-0.59	-0.85	-0.09	0.40	-0.07
M	-0.30	-0.42	-0.30	-0.60	-0.28	1.10	-0.19	-0.21	0.02	-0.15	-0.23	-0.02	-0.34	-0.14	-0.17	-0.12	-0.45	-0.39	-0.03	-0.21
A	-0.30	-0.06	0.03	-0.04	0.03	-0.19	0.08	0.10	-0.10	-0.13	0.18	-0.40	-0.16	-0.26	-0.61	-0.51	-0.40	-0.05	-0.26	-0.07
G	-0.20	-0.12	-0.32	0.21	0.14	-0.21	0.10	-0.19	0.16	-0.02	0.27	-0.01	-0.15	0.02	-0.03	-0.01	-0.23	-0.64	-0.14	0.02
T	0.10	0.15	-0.14	-0.08	-0.57	0.02	-0.10	0.16	0.21	-0.21	-0.54	-0.41	-0.45	0.05	-0.39	-0.50	-0.09	-0.47	-0.07	-0.09
S	-0.24	-0.22	-0.18	-0.33	-0.36	-0.15	-0.13	-0.02	-0.21	0.54	-0.35	0.16	-0.16	-0.36	-0.22	0.13	-0.10	-0.46	-0.23	-0.09
W	-0.18	-0.05	-0.57	0.78	0.17	-0.23	0.18	0.27	-0.54	-0.35	0.90	0.01	-0.45	-0.07	0.33	-0.37	0.14	-0.30	0.25	-0.06
Y	-0.27	-0.55	-0.20	-0.30	-0.34	-0.02	-0.40	-0.01	-0.41	0.16	0.01	0.66	0.20	0.39	-0.40	0.11	0.15	0.05	-0.31	-0.05
P	-0.09	-0.32	0.12	-0.29	-0.59	-0.34	-0.16	-0.15	-0.45	-0.16	-0.45	0.20	0.32	0.25	-0.15	0.34	-0.30	0.19	-0.45	-0.06
H	-0.25	-0.06	-0.38	-0.01	0.23	-0.14	-0.26	0.02	0.05	-0.36	-0.07	0.39	0.25	0.18	-0.28	-0.48	0.06	0.37	-0.02	-0.19
E	-0.14	-0.14	0.04	0.08	-0.20	-0.17	-0.61	-0.03	-0.39	-0.22	0.33	-0.40	-0.15	-0.28	0.39	-0.16	-0.10	0.08	-0.13	-0.12
Q	-0.16	-0.32	0.09	0.08	-0.59	-0.12	-0.51	-0.01	-0.50	0.13	-0.37	0.11	0.34	-0.48	-0.16	0.48	0.04	0.07	-0.64	-0.02
D	-0.22	-0.41	0.08	-0.05	-0.85	-0.45	-0.40	-0.23	-0.09	-0.10	0.14	0.15	-0.30	0.06	-0.10	0.04	0.76	-0.35	-0.44	-0.01
N	-0.08	-0.23	-0.15	-0.34	-0.09	-0.39	-0.05	-0.64	-0.47	-0.46	-0.30	0.05	0.19	0.37	0.08	0.07	-0.35	0.37	0.31	-0.31
K	-0.07	-0.18	0.02	-0.23	0.40	-0.03	-0.26	-0.14	-0.07	-0.23	0.25	-0.31	-0.45	-0.02	-0.13	-0.64	-0.44	0.31	1.59	-0.24
R	-0.32	-0.17	-0.45	-0.01	-0.07	-0.21	-0.07	0.02	-0.09	-0.09	-0.06	-0.05	-0.06	-0.19	-0.12	-0.02	-0.01	-0.31	-0.24	0.53

Supplementary Table 3.3. Symmetric amino acid pairing preference table for type *W* co-evolving residues.

-	I	V	L	F	C	M	A	G	T	S	W	Y	P	H	E	Q	D	N	K	R
I	-0.91	-0.01	0.48	-0.44	0.06	-0.44	-0.31	-0.16	0.06	0.10	-0.57	-0.25	-0.59	0.13	-0.24	-0.49	0.03	-0.35	0.18	-0.19
V	-0.01	-0.91	0.58	0.28	-0.16	0.23	-0.48	-0.12	0.67	0.32	0.16	-0.11	0.10	0.78	0.06	0.34	0.18	-0.41	0.22	-0.03
L	0.48	0.58	-1.34	-0.73	0.16	-0.37	0.10	-0.12	0.52	0.43	0.09	0.15	0.14	-0.49	-0.30	0.13	0.89	-0.38	0.06	-0.48
F	-0.44	0.28	-0.73	-1.11	0.63	-0.87	0.05	0.29	-0.24	0.25	-0.01	-0.35	-0.28	-1.67	0.22	0.58	-0.02	-0.64	-0.20	-0.18
C	0.06	-0.16	0.16	0.63	-1.15	-0.72	-1.17	-0.27	-0.15	-0.16	-7.64	-0.33	-0.17	-1.96	-0.92	0.17	-0.08	-0.04	0.98	-0.66
M	-0.44	0.23	-0.37	-0.87	-0.72	0.34	-0.01	0.45	0.01	0.22	0.59	0.18	-0.62	0.11	-0.58	-0.06	-0.07	-0.87	0.05	-0.08
A	-0.31	-0.48	0.10	0.05	-1.17	-0.01	-0.75	0.51	-0.12	0.14	0.72	-0.41	-0.44	-0.55	-1.07	-0.53	-0.06	-0.65	-0.12	-0.02
G	-0.16	-0.12	-0.12	0.29	-0.27	0.45	0.51	-1.55	0.37	0.46	0.53	0.14	0.13	0.56	0.24	0.61	-0.03	-0.99	0.07	0.22
T	0.06	0.67	0.52	-0.24	-0.15	0.01	-0.12	0.37	-0.90	-0.42	-0.57	-0.11	-0.23	-0.40	-0.51	-0.48	0.88	-0.21	-0.07	0.03
S	0.10	0.32	0.43	0.25	-0.16	0.22	0.14	0.46	-0.42	-1.29	-0.31	-0.07	0.20	-0.85	-0.20	0.52	0.67	-1.04	-0.12	0.01
W	-0.57	0.16	0.09	-0.01	-7.64	0.59	0.72	0.53	-0.57	-0.31	-8.64	0.25	-0.58	0.18	0.39	-1.55	0.08	-1.04	-0.60	-0.66
Y	-0.25	-0.11	0.15	-0.35	-0.33	0.18	-0.41	0.14	-0.11	-0.07	0.25	-3.06	0.71	-0.16	-0.10	1.04	0.85	-1.05	-0.04	0.24
P	-0.59	0.10	0.14	-0.28	-0.17	-0.62	-0.44	0.13	-0.23	0.20	-0.58	0.71	-0.60	0.17	0.23	0.79	0.22	-0.47	0.12	-0.17
H	0.13	0.78	-0.49	-1.67	-1.96	0.11	-0.55	0.56	-0.40	-0.85	0.18	-0.16	0.17	-0.49	0.24	-0.80	-0.59	0.97	0.07	0.09
E	-0.24	0.06	-0.30	0.22	-0.92	-0.58	-1.07	0.24	-0.51	-0.20	0.39	-0.10	0.23	0.24	-0.83	0.19	0.20	0.90	0.21	-0.14
Q	-0.49	0.34	0.13	0.58	0.17	-0.06	-0.53	0.61	-0.48	0.52	-1.55	1.04	0.79	-0.80	0.19	-1.06	0.56	0.62	-0.27	-1.23
D	0.03	0.18	0.89	-0.02	-0.08	-0.07	-0.06	-0.03	0.88	0.67	0.08	0.85	0.22	-0.59	0.20	0.56	-1.43	0.25	-0.21	-0.37
N	-0.35	-0.41	-0.38	-0.64	-0.04	-0.87	-0.65	-0.99	-0.21	-1.04	-1.04	-1.05	-0.47	0.97	0.90	0.62	0.25	-0.08	0.34	0.19
K	0.18	0.22	0.06	-0.20	0.98	0.05	-0.12	0.07	-0.07	-0.12	-0.60	-0.04	0.12	0.07	0.21	-0.27	-0.21	0.34	-0.76	-0.19
R	-0.19	-0.03	-0.48	-0.18	-0.66	-0.08	-0.02	0.22	0.03	0.01	-0.66	0.24	-0.17	0.09	-0.14	-1.23	-0.37	0.19	-0.19	-1.09

Supplementary Table 3.4. Amino acid pairing preference table for type *B* co-evolving residues.

-	I	V	L	F	C	M	A	G	T	S	W	Y	P	H	E	Q	D	N	K	R
I	1.05	-0.05	-0.32	0.21	0.24	-0.82	-1.06	-0.50	-0.22	-0.28	-0.08	-0.63	0.08	-1.36	0.28	0.07	-0.24	-0.74	0.00	-0.13
V	-0.05	0.96	-0.41	-0.16	0.25	-0.62	-0.52	-0.79	-0.36	-0.27	0.05	-0.62	-0.77	-1.07	0.10	-0.24	-0.31	-0.50	-0.25	-0.13
L	-0.32	-0.41	0.55	-0.16	0.39	-0.69	-0.26	-1.17	-0.57	-0.58	-0.90	-0.11	-0.02	-1.23	-0.09	0.49	-0.11	-0.76	-0.28	-0.32
F	0.21	-0.16	-0.16	1.55	-0.30	-0.14	-0.72	0.01	-0.44	-1.17	1.50	-0.06	-0.41	-1.20	-0.31	-0.19	0.09	-0.98	-0.38	-0.56
C	0.24	0.25	0.39	-0.30	2.98	0.61	0.06	0.29	-0.89	-1.47	-0.53	-1.87	-0.13	-0.15	0.31	-1.02	-0.63	-1.28	0.48	-2.37
M	-0.82	-0.62	-0.69	-0.14	0.61	1.83	-0.67	-1.07	-0.35	-0.63	0.10	-0.52	-0.49	-1.09	-0.25	0.27	-0.84	-0.91	0.11	-0.10
A	-1.06	-0.52	-0.26	-0.72	0.06	-0.67	0.67	-0.02	-0.57	-0.41	-0.16	-1.51	-0.46	-1.32	-1.13	-0.53	-0.72	-1.05	-0.54	-0.35
G	-0.50	-0.79	-1.17	0.01	0.29	-1.07	-0.02	0.64	-0.31	-0.53	-1.21	-0.87	-0.72	-1.65	-0.02	-0.80	-0.58	-0.72	-0.26	-0.44
T	-0.22	-0.36	-0.57	-0.44	-0.89	-0.35	-0.57	-0.31	1.06	-0.70	-0.41	-0.54	-0.89	-1.43	-0.11	-0.32	-0.29	-0.94	0.08	0.05
S	-0.28	-0.27	-0.58	-1.17	-1.47	-0.63	-0.41	-0.53	-0.70	1.34	-0.19	-0.27	0.05	-1.60	-0.16	0.15	-0.29	-1.29	-0.18	0.13
W	-0.08	0.05	-0.90	1.50	-0.53	0.10	-0.16	-1.21	-0.41	-0.19	1.12	0.76	-0.06	-1.07	0.38	-0.54	0.27	-1.21	0.55	-0.58
Y	-0.63	-0.62	-0.11	-0.06	-1.87	-0.52	-1.51	-0.87	-0.54	-0.27	0.76	1.66	-0.82	-1.42	-0.47	0.27	0.61	-0.98	-0.65	-1.09
P	0.08	-0.77	-0.02	-0.41	-0.13	-0.49	-0.46	-0.72	-0.89	0.05	-0.06	-0.82	0.91	-0.93	0.11	0.67	-0.16	-1.21	-0.74	0.08
H	-1.36	-1.07	-1.23	-1.20	-0.15	-1.09	-1.32	-1.65	-1.43	-1.60	-1.07	-1.42	-0.93	0.58	-1.11	-1.56	-0.88	0.11	-1.05	-1.51
E	0.28	0.10	-0.09	-0.31	0.31	-0.25	-1.13	-0.02	-0.11	-0.16	0.38	-0.47	0.11	-1.11	0.93	-0.38	-0.30	0.13	-0.31	-0.21
Q	0.07	-0.24	0.49	-0.19	-1.02	0.27	-0.53	-0.80	-0.32	0.15	-0.54	0.27	0.67	-1.56	-0.38	1.33	-0.06	0.51	-0.52	-0.23
D	-0.24	-0.31	-0.11	0.09	-0.63	-0.84	-0.72	-0.58	-0.29	-0.29	0.27	0.61	-0.16	-0.88	-0.30	-0.06	1.51	-0.24	-0.71	-0.13
N	-0.74	-0.50	-0.76	-0.98	-1.28	-0.91	-1.05	-0.72	-0.94	-1.29	-1.21	-0.98	-1.21	0.11	0.13	0.51	-0.24	0.88	0.95	0.04
K	0.00	-0.25	-0.28	-0.38	0.48	0.11	-0.54	-0.26	0.08	-0.18	0.55	-0.65	-0.74	-1.05	-0.31	-0.52	-0.71	0.95	1.38	-0.39
R	-0.13	-0.13	-0.32	-0.56	-2.37	-0.10	-0.35	-0.44	0.05	0.13	-0.58	-1.09	0.08	-1.51	-0.21	-0.23	-0.13	0.04	-0.39	1.18

Supplementary Table 3.5. Contact scoring matrix for all co-evolving residues using 4 Å as cutoff

-	I	V	L	F	C	M	A	G	T	S	W	Y	P	H	E	Q	D	N	K	R
I	-0.07	0.23	0.69	-0.47	-1.69	-1.59	0.61	-1.52	-0.32	0.23	-5.43	0.00	-1.28	0.99	-0.82	0.61	0.07	-0.01	0.15	0.06
V	0.23	-0.29	-0.06	-0.54	-0.92	0.73	-0.18	-1.04	1.48	0.84	-0.10	-0.87	-0.73	-0.89	-0.45	-0.50	0.31	-0.16	0.18	0.69
L	0.69	-0.06	-1.33	-1.63	-3.20	0.28	-0.15	-2.04	1.73	-0.44	-1.39	-0.70	-1.35	0.19	-0.75	-1.19	-0.45	0.98	-0.44	1.45
F	-0.47	-0.54	-1.63	-2.74	-0.53	-0.69	-0.54	-0.26	-2.17	-0.87	-1.77	-1.49	0.33	-0.82	-0.07	-0.18	-0.76	-0.21	0.45	-0.35
C	-1.69	-0.92	-3.20	-0.53	-7.17	-1.46	-1.58	-1.98	-1.44	-4.94	-0.05	-1.15	-4.98	0.31	-5.24	1.98	-0.29	-0.15	0.59	2.28
M	-1.59	0.73	0.28	-0.69	-1.46	0.12	0.02	-0.68	0.26	-0.51	-1.33	-1.12	1.53	0.34	0.30	-0.86	-5.10	1.11	-1.41	-1.61
A	0.61	-0.18	-0.15	-0.54	-1.58	0.02	0.23	0.03	0.98	-0.52	-1.39	-1.80	-0.86	-0.42	-0.53	1.17	-0.08	1.10	0.46	-0.46
G	-1.52	-1.04	-2.04	-0.26	-1.98	-0.68	0.03	-1.47	-0.51	1.47	-5.15	-2.85	0.29	0.13	-1.95	2.02	-0.32	-0.71	0.45	0.03
T	-0.32	1.48	1.73	-2.17	-1.44	0.26	0.98	-0.51	-1.59	-0.15	0.35	-1.24	-1.54	-0.28	-1.60	0.67	-2.15	-0.71	-0.47	-0.63
S	0.23	0.84	-0.44	-0.87	-4.94	-0.51	-0.52	1.47	-0.15	-2.07	0.98	-1.87	-1.76	1.32	0.87	1.18	-1.11	0.59	1.75	-0.09
W	-5.43	-0.10	-1.39	-1.77	-0.05	-1.33	-1.39	-5.15	0.35	0.98	-6.65	-0.38	-5.07	-0.86	-1.71	-4.60	-0.08	1.13	-1.31	0.13
Y	0.00	-0.87	-0.70	-1.49	-1.15	-1.12	-1.80	-2.85	-1.24	-1.87	-0.38	-2.43	-1.76	-0.31	2.40	0.65	1.32	0.03	1.06	-0.73
P	-1.28	-0.73	-1.35	0.33	-4.98	1.53	-0.86	0.29	-1.54	-1.76	-5.07	-1.76	-3.33	-1.12	1.20	-0.37	-0.56	-1.29	-0.92	-1.25
H	0.99	-0.89	0.19	-0.82	0.31	0.34	-0.42	0.13	-0.28	1.32	-0.86	-0.31	-1.12	-3.33	0.52	-0.54	1.41	2.37	-1.20	-0.20
E	-0.82	-0.45	-0.75	-0.07	-5.24	0.30	-0.53	-1.95	-1.60	0.87	-1.71	2.40	1.20	0.52	-0.28	-1.55	-0.54	-2.27	0.90	-0.16
Q	0.61	-0.50	-1.19	-0.18	1.98	-0.86	1.17	2.02	0.67	1.18	-4.60	0.65	-0.37	-0.54	-1.55	-2.71	-0.67	-2.10	-1.09	-1.45
D	0.07	0.31	-0.45	-0.76	-0.29	-5.10	-0.08	-0.32	-2.15	-1.11	-0.08	1.32	-0.56	1.41	-0.54	-0.67	-2.40	0.48	0.52	0.26
N	-0.01	-0.16	0.98	-0.21	-0.15	1.11	1.10	-0.71	-0.71	0.59	1.13	0.03	-1.29	2.37	-2.27	-2.10	0.48	-2.55	-1.09	-0.45
K	0.15	0.18	-0.44	0.45	0.59	-1.41	0.46	0.45	-0.47	1.75	-1.31	1.06	-0.92	-1.20	0.90	-1.09	0.52	-1.09	-1.15	-1.04
R	0.06	0.69	1.45	-0.35	2.28	-1.61	-0.46	0.03	-0.63	-0.09	0.13	-0.73	-1.25	-0.20	-0.16	-1.45	0.26	-0.45	-1.04	-1.42

Supplementary Table 3.6. Contact scoring matrix for all co-evolving residues using 5 Å as cutoff

-	I	V	L	F	C	M	A	G	T	S	W	Y	P	H	E	Q	D	N	K	R
I	0.19	0.67	0.73	-0.73	-2.21	-0.78	0.53	0.48	-0.34	-0.26	-5.43	-0.25	-1.53	0.48	-1.24	0.31	0.30	-0.04	-0.16	-0.40
V	0.67	-0.55	0.13	0.30	-1.14	0.45	-0.23	-0.60	1.35	0.64	-0.01	-1.12	-0.67	-1.15	-0.80	-0.71	0.55	-0.04	-0.05	0.34
L	0.73	0.13	-1.18	-0.31	-2.90	0.17	0.14	-2.06	1.68	-0.59	-1.63	-0.73	-1.06	-0.11	-0.89	-1.02	0.11	0.84	0.44	1.14
F	-0.73	0.30	-0.31	-1.16	-0.85	-1.03	0.00	-0.03	-1.63	-1.52	-1.56	-1.84	-0.04	-1.72	-0.73	-0.59	-1.19	-0.26	0.19	-0.90
C	-2.21	-1.14	-2.90	-0.85	-7.17	-1.62	-1.24	-2.06	-1.17	-1.56	0.34	-0.93	-1.01	-0.16	0.48	1.77	-1.01	-0.43	0.35	2.02
M	-0.78	0.45	0.17	-1.03	-1.62	0.15	-0.39	-1.38	-0.15	-1.41	-1.36	-1.54	1.03	-0.51	-0.59	-1.35	-2.49	1.14	-2.09	-2.66
A	0.53	-0.23	0.14	0.00	-1.24	-0.39	0.29	-0.17	1.02	-0.78	-1.31	-1.77	-0.69	-0.61	-1.06	1.16	-0.02	0.94	0.34	-0.53
G	0.48	-0.60	-2.06	-0.03	-2.06	-1.38	-0.17	-0.04	-0.22	1.06	-1.01	0.13	-0.13	-0.42	-1.54	1.50	-0.15	0.17	0.41	-0.20
T	-0.34	1.35	1.68	-1.63	-1.17	-0.15	1.02	-0.22	-1.35	-0.36	0.36	-1.26	-1.27	-0.70	-1.28	0.50	-1.71	-0.80	-0.68	-0.82
S	-0.26	0.64	-0.59	-1.52	-1.56	-1.41	-0.78	1.06	-0.36	-0.50	1.04	-1.30	-2.17	0.64	-0.14	1.53	-0.70	0.08	1.12	-0.39
W	-5.43	-0.01	-1.63	-1.56	0.34	-1.36	-1.31	-1.01	0.36	1.04	-6.64	-0.31	-5.06	-0.80	-1.90	-4.60	-0.39	1.11	-1.05	-0.40
Y	-0.25	-1.12	-0.73	-1.84	-0.93	-1.54	-1.77	0.13	-1.26	-1.30	-0.31	-2.57	-1.20	-0.87	1.91	1.30	0.97	0.81	0.79	-1.13
P	-1.53	-0.67	-1.06	-0.04	-1.01	1.03	-0.69	-0.13	-1.27	-2.17	-5.06	-1.20	-2.97	-1.60	0.59	-0.48	1.21	1.11	-0.19	-0.48
H	0.48	-1.15	-0.11	-1.72	-0.16	-0.51	-0.61	-0.42	-0.70	0.64	-0.80	-0.87	-1.60	-2.19	1.03	-0.89	0.77	1.83	-0.56	-0.89
E	-1.24	-0.80	-0.89	-0.73	0.48	-0.59	-1.06	-1.54	-1.28	-0.14	-1.90	1.91	0.59	1.03	-0.48	0.85	0.16	-3.11	0.39	-0.22
Q	0.31	-0.71	-1.02	-0.59	1.77	-1.35	1.16	1.50	0.50	1.53	-4.60	1.30	-0.48	-0.89	0.85	-2.54	-0.89	-2.29	-1.25	-1.52
D	0.30	0.55	0.11	-1.19	-1.01	-2.49	-0.02	-0.15	-1.71	-0.70	-0.39	0.97	1.21	0.77	0.16	-0.89	-0.74	0.00	0.11	-0.32
N	-0.04	-0.04	0.84	-0.26	-0.43	1.14	0.94	0.17	-0.80	0.08	1.11	0.81	1.11	1.83	-3.11	-2.29	0.00	-1.88	-1.21	1.06
K	-0.16	-0.05	0.44	0.19	0.35	-2.09	0.34	0.41	-0.68	1.12	-1.05	0.79	-0.19	-0.56	0.39	-1.25	0.11	-1.21	-0.39	-0.86
R	-0.40	0.34	1.14	-0.90	2.02	-2.66	-0.53	-0.20	-0.82	-0.39	-0.40	-1.13	-0.48	-0.89	-0.22	-1.52	-0.32	1.06	-0.86	-1.77

Supplementary Table 3.7. Contact scoring matrix for all co-evolving residues using 6 Å as cutoff

-	I	V	L	F	C	M	A	G	T	S	W	Y	P	H	E	Q	D	N	K	R
I	0.10	0.66	0.62	-0.86	0.21	-0.84	0.42	0.37	-0.44	-0.49	-5.42	-0.35	-1.46	0.47	-1.26	0.27	0.26	-0.16	-0.39	-0.67
V	0.66	-0.64	0.18	0.19	-1.09	0.61	-0.30	-0.67	1.25	0.66	0.09	-0.91	-0.61	-1.02	-0.79	-0.19	0.41	-0.18	-0.26	0.04
L	0.62	0.18	-1.29	-0.48	-2.87	0.13	0.26	-2.05	1.57	-0.78	-1.54	-0.85	-1.10	0.07	-0.92	-1.07	0.10	0.85	1.18	0.84
F	-0.86	0.19	-0.48	-1.37	-0.88	-1.06	-0.15	0.16	-0.11	-1.51	-1.54	-1.84	0.15	-1.66	-0.53	-0.65	-0.77	-0.41	0.26	-0.89
C	0.21	-1.09	-2.87	-0.88	-7.17	-1.56	-1.20	-0.44	-1.11	-1.68	0.45	-0.84	-0.93	0.00	0.53	1.72	0.55	-0.48	0.05	1.72
M	-0.84	0.61	0.13	-1.06	-1.56	0.17	-0.50	-1.35	-0.28	-1.54	-1.25	-1.74	0.94	-0.48	-0.68	-1.39	-1.48	0.84	-2.15	-2.78
A	0.42	-0.30	0.26	-0.15	-1.20	-0.50	0.17	-0.28	0.92	-0.97	-1.22	-1.71	-0.70	-0.66	-1.05	1.11	0.42	0.75	0.24	0.36
G	0.37	-0.67	-2.05	0.16	-0.44	-1.35	-0.28	-0.12	-0.29	1.15	-0.25	0.63	0.15	-0.38	-1.56	1.43	0.19	0.09	0.17	-0.52
T	-0.44	1.25	1.57	-0.11	-1.11	-0.28	0.92	-0.29	-1.45	-0.49	0.29	-0.17	-1.21	-0.55	-1.30	0.63	-1.27	1.28	-0.92	-0.77
S	-0.49	0.66	-0.78	-1.51	-1.68	-1.54	-0.97	1.15	-0.49	-0.74	0.76	-1.59	0.32	0.46	-0.30	1.29	0.56	0.21	1.46	1.28
W	-5.42	0.09	-1.54	-1.54	0.45	-1.25	-1.22	-0.25	0.29	0.76	-6.64	-0.17	-5.06	-0.59	-1.04	-4.59	-0.38	1.10	-1.05	-0.48
Y	-0.35	-0.91	-0.85	-1.84	-0.84	-1.74	-1.71	0.63	-0.17	-1.59	-0.17	-2.46	-0.89	-0.70	1.93	1.29	0.82	0.66	0.66	-1.24
P	-1.46	-0.61	-1.10	0.15	-0.93	0.94	-0.70	0.15	-1.21	0.32	-5.06	-0.89	-2.91	-1.43	0.60	-0.56	1.07	1.07	-0.37	-0.59
H	0.47	-1.02	0.07	-1.66	0.00	-0.48	-0.66	-0.38	-0.55	0.46	-0.59	-0.70	-1.43	-2.17	1.15	-0.71	0.72	1.72	-0.68	-0.92
E	-1.26	-0.79	-0.92	-0.53	0.53	-0.68	-1.05	-1.56	-1.30	-0.30	-1.04	1.93	0.60	1.15	-0.44	0.88	0.05	-3.10	0.26	-0.23
Q	0.27	-0.19	-1.07	-0.65	1.72	-1.39	1.11	1.43	0.63	1.29	-4.59	1.29	-0.56	-0.71	0.88	-2.32	-0.98	-2.32	-1.28	-1.62
D	0.26	0.41	0.10	-0.77	0.55	-1.48	0.42	0.19	-1.27	0.56	-0.38	0.82	1.07	0.72	0.05	-0.98	-1.02	0.51	-0.13	-0.68
N	-0.16	-0.18	0.85	-0.41	-0.48	0.84	0.75	0.09	1.28	0.21	1.10	0.66	1.07	1.72	-3.10	-2.32	0.51	-1.91	-1.51	0.70
K	-0.39	-0.26	1.18	0.26	0.05	-2.15	0.24	0.17	-0.92	1.46	-1.05	0.66	-0.37	-0.68	0.26	-1.28	-0.13	-1.51	-0.55	0.33
R	-0.67	0.04	0.84	-0.89	1.72	-2.78	0.36	-0.52	-0.77	1.28	-0.48	-1.24	-0.59	-0.92	-0.23	-1.62	-0.68	0.70	0.33	-2.09

Supplementary Table 3.8. Contact scoring matrix for all co-evolving residues using 7 Å as cutoff

-	I	V	L	F	C	M	A	G	T	S	W	Y	P	H	E	Q	D	N	K	R
I	-0.01	0.42	0.39	0.04	0.09	-0.56	0.18	-0.17	-0.32	-0.16	-5.61	-0.78	-1.16	0.06	0.54	0.10	0.12	-0.22	-0.27	-0.27
V	0.42	-0.74	0.10	-0.04	-1.58	0.47	-0.37	-0.90	1.00	0.50	-0.30	-1.00	-0.63	-1.04	-0.16	-0.13	0.38	-0.21	-0.20	-0.05
L	0.39	0.10	-1.11	0.50	-2.69	0.48	0.17	-1.37	1.28	-0.81	-1.66	-0.89	-1.23	0.92	-0.78	-0.72	0.08	1.17	1.14	0.63
F	0.04	-0.04	0.50	-1.91	-0.93	-0.74	-0.62	-0.19	-0.45	-1.44	-6.21	-1.37	0.32	-1.80	-0.71	0.53	-0.77	-0.90	0.51	-1.27
C	0.09	-1.58	-2.69	-0.93	-7.13	-1.55	-1.88	-0.79	0.44	-4.93	-5.34	-4.56	-4.94	-1.17	0.37	1.66	0.25	-0.30	0.07	1.73
M	-0.56	0.47	0.48	-0.74	-1.55	0.18	-0.71	-0.30	-0.55	-1.11	-1.36	-1.76	0.80	-0.88	-0.20	-0.54	-1.69	1.12	-2.58	-1.65
A	0.18	-0.37	0.17	-0.62	-1.88	-0.71	0.02	0.07	0.70	-0.84	-0.46	-1.31	0.73	-0.80	-0.24	1.04	0.21	0.77	0.24	0.11
G	-0.17	-0.90	-1.37	-0.19	-0.79	-0.30	0.07	-0.64	-0.88	0.78	1.37	0.41	1.10	-1.05	-0.21	1.19	0.29	0.09	0.40	-0.45
T	-0.32	1.00	1.28	-0.45	0.44	-0.55	0.70	-0.88	-1.47	-0.66	0.52	-0.40	-1.45	1.41	-1.47	0.49	0.41	1.20	-0.84	-0.93
S	-0.16	0.50	-0.81	-1.44	-4.93	-1.11	-0.84	0.78	-0.66	-0.93	0.88	-1.85	0.16	0.09	-0.44	1.09	0.39	1.19	1.35	1.06
W	-5.61	-0.30	-1.66	-6.21	-5.34	-1.36	-0.46	1.37	0.52	0.88	-7.02	-0.58	-5.23	-4.77	0.06	-0.03	-0.24	1.26	-4.71	-0.26
Y	-0.78	-1.00	-0.89	-1.37	-4.56	-1.76	-1.31	0.41	-0.40	-1.85	-0.58	-2.38	-1.12	-1.42	1.60	1.48	0.70	0.40	0.59	-1.27
P	-1.16	-0.63	-1.23	0.32	-4.94	0.80	0.73	1.10	-1.45	0.16	-5.23	-1.12	-2.88	-2.25	0.43	-0.70	0.90	0.86	-0.40	-0.96
H	0.06	-1.04	0.92	-1.80	-1.17	-0.88	-0.80	-1.05	1.41	0.09	-4.77	-1.42	-2.25	-2.85	1.69	-1.31	0.39	1.25	-0.86	-1.19
E	0.54	-0.16	-0.78	-0.71	0.37	-0.20	-0.24	-0.21	-1.47	-0.44	0.06	1.60	0.43	1.69	-0.87	0.63	-0.11	-3.16	0.12	-0.55
Q	0.10	-0.13	-0.72	0.53	1.66	-0.54	1.04	1.19	0.49	1.09	-0.03	1.48	-0.70	-1.31	0.63	-2.15	-1.07	-2.47	-1.22	-1.45
D	0.12	0.38	0.08	-0.77	0.25	-1.69	0.21	0.29	0.41	0.39	-0.24	0.70	0.90	0.39	-0.11	-1.07	-1.08	0.33	0.16	-0.85
N	-0.22	-0.21	1.17	-0.90	-0.30	1.12	0.77	0.09	1.20	1.19	1.26	0.40	0.86	1.25	-3.16	-2.47	0.33	-1.80	-1.27	0.41
K	-0.27	-0.20	1.14	0.51	0.07	-2.58	0.24	0.40	-0.84	1.35	-4.71	0.59	-0.40	-0.86	0.12	-1.22	0.16	-1.27	-0.51	0.34
R	-0.27	-0.05	0.63	-1.27	1.73	-1.65	0.11	-0.45	-0.93	1.06	-0.26	-1.27	-0.96	-1.19	-0.55	-1.45	-0.85	0.41	0.34	-1.32

CHAPTER 4

APPLICATION OF CO-EVOLUTION AND RESIDUE CONTACT PREDICTION METHODS TO SEPTINS[3]

[3] Fangfang Pan, Dongsheng Che, Michelle Momany, Liming Cai, and Russell L. Malmberg. To be submitted to *Proteins:structure, function, bioinformatics*

ABSTRACT

Septins form filaments, act as a barrier, recruit proteins and play key roles in cell division. We studied the co-evolution of members of the septin gene family. We focused on the co-evolution of 5 septins, Cdc3, Cdc10, Cdc11, Cdc12 and Shs1 that interact to form a filamentous structure at the mother-bud neck in *Saccharomyces cerevisiae*. We looked at the co-evolution among these five subunits as well as within each subunit. Using mutual information and a contact prediction matrix, we identified some sites on septins that were co-evolving and interacting. We also tested the co-evolution of each of these five septins with formin (Bni1) and myosin (Myo1). We were able to suggest the septin subunit that interacted with formin and myosin. This provided good candidates for the study of septin function.

INTRODUCTION

All septins can be placed in one of 5 groups, CDC3, CDC10, CDC11, CDC12 or ASPE, names adopted from well-studied septins in *Saccharomyces cerevisiae* and *Aspergillus nidullans* [3]. On the amino acid sequence level, all septins have a conserved GTPase domain, a variable N-terminus and a variable C-terminus. Some septins have a predicted coiled-coil structure at the C-terminus. The coiled-coil is a protein structural motif, where several alpha-helixes are coiled together, and many of the coiled-coil structures mediate protein-protein interactions [74]. CDC10 type septins lack this coiled-coil region. In addition to these well-defined regions, polybasic residues thought to bind phosphoinositide can often be found preceding the GTPase domain. A conserved "septin unique region" of unknown function can be found following the GTPase domain [4, 59].

In both fungi and animals, septin monomers assemble into higher order structures [4, 56]. Septins have been most studied in *S. cerevisiae*, which has seven septins Cdc3, Cdc10, Cdc11,

Cdc12, Shs1, Spr3 and Spr28. Among those, Cdc3, Cdc10, Cdc11, Cdc12 and Shs1 proteins are structural components of the septin complex and are known to interact with one another to form the parallel filaments often observed at the mother-bud neck of yeast [4].

A septin complex structure model was proposed, based on experiments in *S. cerevisiae*[4]. This model indicates possible septin subunit arrangements in the septin heteropentamer complex. Septin proteins first interact with a septin protein of the same type, and form dimers. Cdc3, Cdc11 and Cdc12 dimers further interact with each other and form the Cdc3-Cdc12-Cdc11 structure. Cdc12 interacts with Cdc3 at the C-terminus and with Cdc11 at the N-terminus. The existence of Cdc11 doesn't affect the interaction of Cdc12 with Cdc3. Cdc11 interacts with Cdc12 in the presence of Cdc3. These Cdc3-Cdc12-Cdc11 units are repeated in the filament. Cdc10 interacts with an interface created by the interaction of Cdc3 and Cdc12 and works as a bridge to connect these units. Shs1 is peripheral. Shs1 interacts with Cdc11 in the filament structure and is not essential for the viability of yeast cells. The interaction of Cdc11 and Shs1 is also through the C-terminus.

As most septins exhibit important functions through the hetero-oligomer filament structures, it is reasonable to postulate co-evolution between septin units. Experiments have been performed to identify sequence regions important for the interactions; however the results do not always agree with each other [4, 102]. We would like to add more clues to the septin interactions from a different angle by looking at co-evolving residues that may be physically close to each other.

Two protein families, the myosins and formins, are also important in cytokinesis, and are functionally related to septins. However details of their interactions with septins are still under investigation. Myosins forma large super family of actin-based motor proteins found in eukaryotes [103, 104]. The best studied myosin is myosin class II. It can be found in all

eukaryotes except for plants [105]. The typical myosin type II heavy chain sequence contains a globular head domain at the N-terminus, and a tail domain at the C-terminus, which can form a coiled-coil, dimerize and further polymerize into filaments [105, 106]. The head and tail domains are connected by a neck domain, which can also bind the myosin light chain. Co-evolution within the myosin sequence was previously suggested by identical or similar phylogenetic patterns of this super family [106].

Myo1p [gi|6321812] is a type II myosin heavy chain in *S. cerevisiae*. It plays critical role in cytokinesis [107]. Myo1 localizes to the actomyosin contractile ring at the bud site and remains at the mother-bud neck [108]. The septin ring is required for this contractile ring formation at the bud neck [108-110]. A mutation in septin Cdc12 will abolish the localization of myosin [110]. Inter-dependence of the head, neck, and tail domains of myosin have been suggested [106]. We are interested to know whether there is evidence of co-evolution and interactions between myosin and septin as these two protein families are functionally closely related.

Formins are required for cytokinesis and maintenance of cell polarity [111, 112]. A formin protein is typically around 1500 amino acids long. The protein sequence contains two conserved domains FHl and FH2; and FHl are extremely proline rich [112]. There is usually one coiled-coil domain lying N-terminal to the FHl domain, and one within the region between the FH2 domain and the C-terminus. Bni1p [gi|6324058], a formin family protein, has the function of nucleating the assembly of actin filaments [111]. Bni1 was identified by genetic interaction with septin Cdc12 in yeast [113]. Bni1 is required for the assembly of the septin ring during the initiation of budding, but not for maintenance [114]. In the bni1 mutant, septins were recruited to the incipient budding site but not assembled, and septins remained at the polarized growing sites.

Point mutants of Bni1 that are defective in actin cable formation also exhibited septin ring assembly defects.

Here we attempted to computationally identify the co-evolving sites, and presumably interacting regions, among five septins and any of the five septins with myosin and formin. We used the mutual information methods described in Chapter 3 to help narrow down possible interaction sites that could be targets for further mutagenesis analysis.

MATERIALS AND METHODS

We followed the methods developed in Chapter 3 and applied them to five septins, one formin and one myosin sequence in *Saccharomyces cerevisiae*. First, we collected the available homologues, aligned multiple sequences and analyzed mutual information for co-evolution. Second, we used the contact scoring matrix to assign scores to each co-evolving residues and identified those that were likely to be in contact in 3D structure.

Search for homologous sequences

The sequences we used were five septins (Cdc3, gi|6323346; Cdc10, gi|6319847; Cdc11, gi|6322536; Cdc12, gi|6321899; Shs1, gi|6319976), myosin (Myo1p, gi|6321812), and formin (Bni1p, gi|6324058). They were extracted from Genbank and used as queries in BLASTP searches, run against the non-redundant database on NCBI with default parameter settings[95]. BLASTP results with lowest *e*-value in each organism were kept. We further sorted through the sequences and only kept those sequences where corresponding homologs from the same organism (iso-organismal homologs) existed for both proteins studied.

Align sequences and merge alignments of two polymers

ClustalW was used to align the homologous sequences of each protein [96]. Protein alignments of any two proteins were merged to a single file and MI was calculated between and within proteins. We are interested in looking at the pair wise combination among 5 septins, Cdc3, Cdc10, Cdc11, Cdc12 and Shs1, and the combination of any of the five septins with myosin or formin and we are specifically interested in the interactions between those proteins.

Search for co-evolving residues

As mentioned in Pan *et al*. 2007, we only considered the positions that had appropriate sequence variation by using positions with entropy greater than 0.3. Positions are considered co-evolving when the *Z*-score (the units of standard deviation from the mean) was greater than 4. We looked at both type *W* (Within a protein) co-evolving residues and type *B* (Between proteins) co-evolving residues.

Predict interacting co-evolving residues

After using the previous steps to identify co-evolving residues, we predicted the residue contacts by calculating scores from the contact scoring matrix. The average score (*AS*)

was used to estimate whether two positions were in contact. A suitable threshold *T* was chosen and compared with the value of *AS* to determine if the co-evolving positions were in physical contact or not. The Graphviz program (http://www.graphviz.org) was used to represent the results of co-evolution and interactions graphically.

RESULTS

Five septins, Cdc3, Cdc10, Cdc11, Cdc12 and Shs1, that are the components of the septin heterocomplex were analyzed. The co-evolution of the residues both within sequences and

between any two sequences was identified and summarized in Table 4.1. The residue contacts within each septin sequence and between any two septins were studied and summarized in Table 4.2. The results are graphically represented (Figure 4.3-4.8). The predicted interactions between septins and myosin or formin were also studied. Those results are presented in Figure 4.9-Figure 4.11.

As shown in *Pan et al 2007*, different thresholds of the T score can result in different sensitivity and specificity. We used some random sequences that are unlikely to physically interact and calculated their AS score as described in methods. Protein sequences were extracted from the mouse LOCATE subcellular localization database (http://locate.imb.uq.edu.au). 100 proteins sequences were randomly picked with nuclear localization and 100 protein sequences were randomly extracted with plasma membrane localization. The physical interaction score AS was calculated between all of the residue positions in each nucleus sequence with all of the residue positions in each plasma membrane protein. The mean AS score for the AS score distribution of those random sequences was -0.10 and the standard deviation was 0.07. Thus, a T score of 0.18 (-0.10 + 4 * 0.07) was used in our analysis.

Co-evolution and residue interaction prediction within septin sequences

The five septin proteins first form dimers from two identical septins before interacting to form septin filaments [4]. We are not yet able to disginguish interactions within a single septin monomer from interactions that occur between two identical septin subunits in a dimer. There are many co-evolving residues within each of the septin sequences forming co-evolution networks (Figure 4.2). The number of co-evolving pairs with each septin sequence were around 20~30 (Table 4.1). Figure 4.1a shows an example of co-evolving residues between septin Cdc3p and Cdc12p. H329 is in the GTP binding domain after the G4 motif. Q416, S428, K429, F452

are in the variable region near the C-terminus before the predicted coiled-coil. In Cdc12p, I72 is in the GTP binding domain between G1 and G3 motifs (Figure 4.2b). E346, E368 are in are in the variable region near C-terminus before the predicted coiled-coil. In Cdc10, where is no predicted coiled-coil, there are two pairs (Figure 4.2c). I22 is N-terminal to the polybasic region. L69 is in the GTP binding domain between G1 and G3 motifs. R127 is between G3 and G4 motif. K304 is at the C-terminal. In Cdc11, E363, E369, F376 all lie in the region variable region near C-terminus before the predicted coiled-coil. In Shs1, E258 and N310, are both right after the GTPase domain. Y465, Q477 and S486 are all in the predicted coiled-coil domain (Figure 4.2e).

Among those co-evolving pairs, some are predicted to be physically close in 3D structure using a relatively strict criteria (T=0.18) (Figure 4.3, Table 4.2). The number of predicted interaction pairs are around 2~3 within a septin. The location of those interaction pairs are summarized in Table 4.3. For example, in Cdc3p, two pairs of interaction are within the variable region before the predicted coiled-coil domain and one pair is between this domain and the GTP_CDC domain.

The co-evolution and interaction prediction of septins with other septins

There are large networks of co-evolution between any two septin sequences from these five septins. There are totally 1250 co-evolving pairs between any two different septins. Those results are too numerous to be clearly represented graphically in the space provided here. The total number of co-evolving pairs in each pair of sequences is summarized in Table 4.1. Among all those co-evolving residues, 168 pairs are predicted to be physically close in 3D structure (Table 4.2).

The168 interactions pairs between different septin proteins are shown in Figures 4.4~4.8. In the Cdc3 sequence, amino acids from positions 413-520 lie in the C-terminal region and residues from 476-507 are in the predicted coiled-coil region [4]. The majority of the predicted co-

evolving and interacting residues between amino acids in Cdc3 and residues from other septins are within the C-terminal region, and sometimes the GTP_CDC region (Figure 4.4). Figure 4.5 shows an example of the interaction between the C-terminal of Cdc3p with the GTP_CDC domain of Cdc12p. Figure 4.6 shows the interaction of Cdc12p with other septin proteins. The variable C-terminus region is involved in some interactions. The N-terminus and the septin unique region also have some interactions. One interesting residue is E258 C-terminal to the GTP_CDC domain in Shs1 sequence. It is co-evolving and predicted to interact with more than one residue in each of Cdc3p, Cdc12p, Cdc10p and Cdc11p sequences.

According to the results above, we can tell that there are some residues that are worth further study. In Cdc3p, position 429 and position 449 are the ones that have several different interaction partners with different proteins. In Cdc3p, most of the interactions are from the variable C-terminus and sometimes the GTP_CDC domain. Position 72 in Cdc12p, position 300 in Cdc10p are of interest. In Cdc10p, many of the interacting residues are from the septin unique domain. In septin Cdc12p, there are more interactions from the coiled-coil region than occur in the other septin sequences. In Cdc11p, positions 315 and 354 also have more than one interaction pair. Most of the interactions in Cdc11p are from the C-terminus. In Shs1Position 311 forms a one-to-several interaction pattern with residues. Position 259 in addition to 258 stands out in Shs1p as it forms one-to-several pattern with more than one other septins. In Shs1p most of the interactions are from the GTP_CDC domain.

The co-evolution and interaction prediction of septins with formin and myosin

Cdc12p is predicted to have more interactions with formin than the other four septins do, and Shs1 has the fewest predicted co-evolving and interacting residues with formin (Figure 4.9~4.10, Table 4.2). Cdc12p has around 3~4 times of the interactions between Cdc11p, Cdc10p and

Cdc3p with Bni1, and 12 times the number of interaction between Shs1p and Bni1p. Figure 4.11 gives the predicted interactions between these septins and myosin. Some of the positions were previously predicted in septin-septin interactions. Cdc3p has twice as many predicted interactions with myosin as the other four septins. Shs1p has the least interaction numbers with Myo1p. Still, most of the interactions lie in the C-terminus. Surprisingly, it is usually not the predicted coiled-coil region. Many of the interaction sites are at the variable region near the C-terminus. We can see some obvious networks there. For example, the S509 in Cdc3p is predicted to be co-evolving with six Myo1p residues.

For the predicted interactions of formin with septins, we seldom see the involvement of the N-terminus region. Only one position in Cdc10p was predicted to interact with formin (Figure 4.9). In Cdc3p, most of the interaction sites are in the variable C-terminus region. In Cdc10p, most interactions are in the GTP_CDC region. In Cdc11p, most interactions are in the variable C-terminus and the predicted coiled-coil region. In Cdc12p, where there are a lot of predicted interactions with Bni1p, most positions are from the variable C-terminus and some are from the GTP_CDC domain.

For the interaction between myosin and septins, there isn't much residue overlap between that of septin-septin interactions. In Cdc3p, the variable C and GTP_CDC domain are the major interaction regions. In Cdc12p, there are many residues from the N-terminus region that are predicted to interact with myosin. In Cdc10p, the septin unique region and the GTP_CDC domain interact with myosin. In Cdc11p, half of the interactions are from the septin unique region.

DISCUSSION

The application of mutual information analysis to protein sequences were shown to detect co-evolution [90]. The application of mutual information analysis to co-evolution and residue contact prediction can help to detect interesting sites in protein sequences. Though the sensitivity and specificity of this method is still not yet satisfactory, it looks at the interaction from a co-evolution direction in stead of studying conserved positions.

From the analysis of septins with formin and myosin, two septin proteins are likely to co-evolve and interact with them, Cdc12p with Bni1p and Cdc3p with Myo1p. The predicted interaction sites and co-evolving residues are more than the other septins (Table 4.1 and 4.2).

In septins, the variable C-terminus region before the predicted coiled-coil domain is not conserved. However, it shows signs of co-evolution and some positions exhibit strong signals. Though the 3D structure of human septin complex was recently released, the 3D structure of the septin C-terminus is not yet well resolved and it is difficult to make comparisons at this stage between this structure and the co-evolving positions[102].

The results presented here are only part of the application. One can change the parameter T setting according to the application purpose to achieve different sensitivity and specificity (Pan et al., chapter 3). In these experiments, we set a relative high T value (4 standard deviations above the random mean) so that we had lower sensitivity and higher specificity, which is particularly convenient for graphical interpretation of the results. If a residue is not detected here, it doesn't mean that residue is not involved in physical interaction.

We identified some co-evolving interactions among septins as well as between septins and myosin and formin. Cdc12p was more likely to interact with Bni1p and Cdc3p was more likely

to interact with Myo1p. However, since the sensitivity and specificity of these methods are not high, the test results were not yet very different from the control. It may need improvement before being further applied to other proteins.

Figure 4.1. Primary amino acid structures and location of domains of five *S. cerevisiae* septin proteins, Cdc3p, Cdc12p, Cdc10p, Cdc11p and Shs1p. The margins of the domains are adapted from Versele et al, 2004. The sequences are shown in ten-residue blocks. Left hand side numbers are the sequence residue position numbers. The numbers on the top of each alignment is the alignment position. The Clustal consensus line marks the conserved positions. Three vertical bars are the boundaries of domains. “a” is the boundary of N-terminus variable domain and GTP_CDC domain. “b” is the boundary of GTP_CDC doman and the septin unique domain. “c” is the boundary of septin unique domain and the c-terminus domain. Within the c-terminus, some proteins have predicted coiled-coil domain and the coiled-coil is represented as horizontal dashed lines under the alignment.

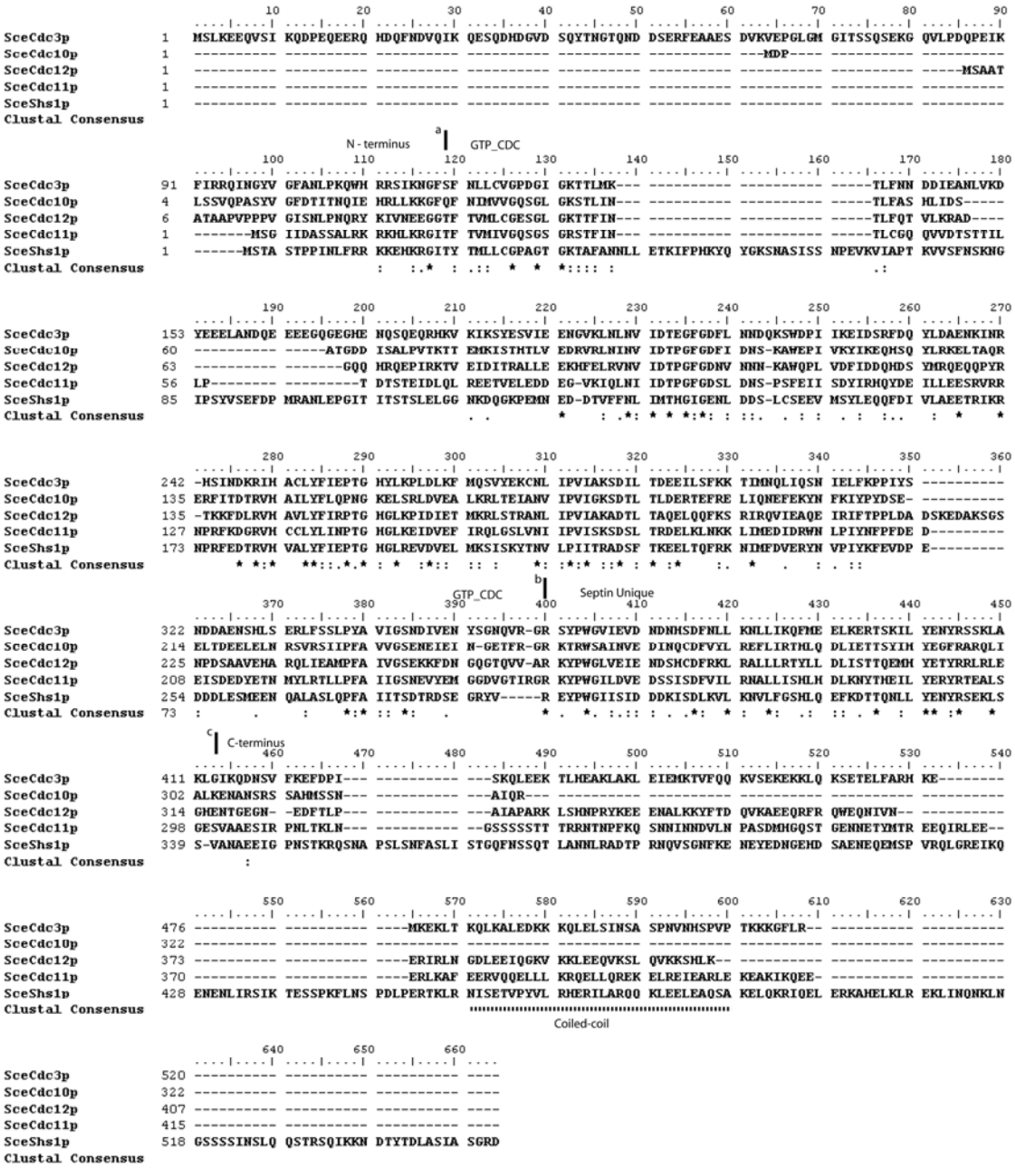
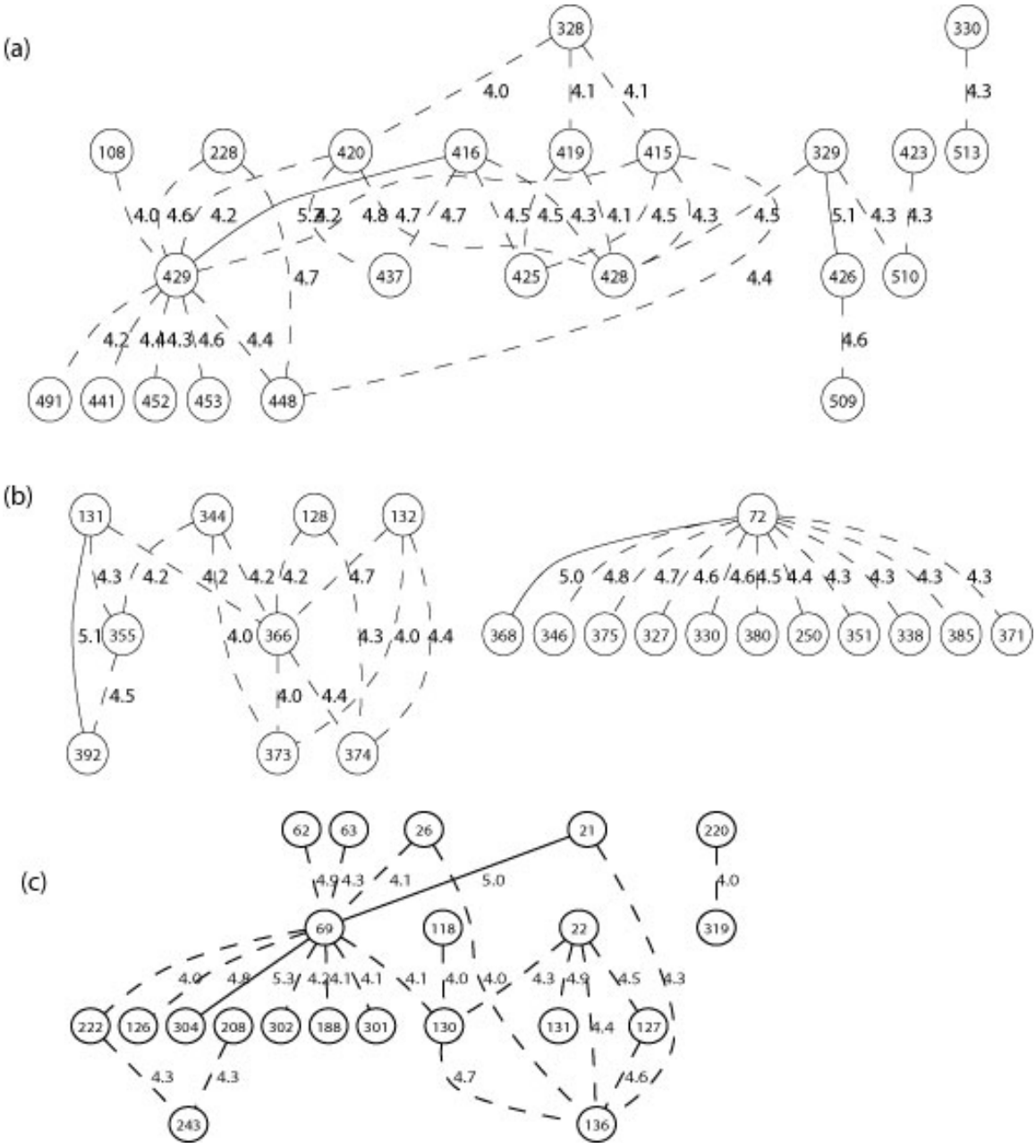
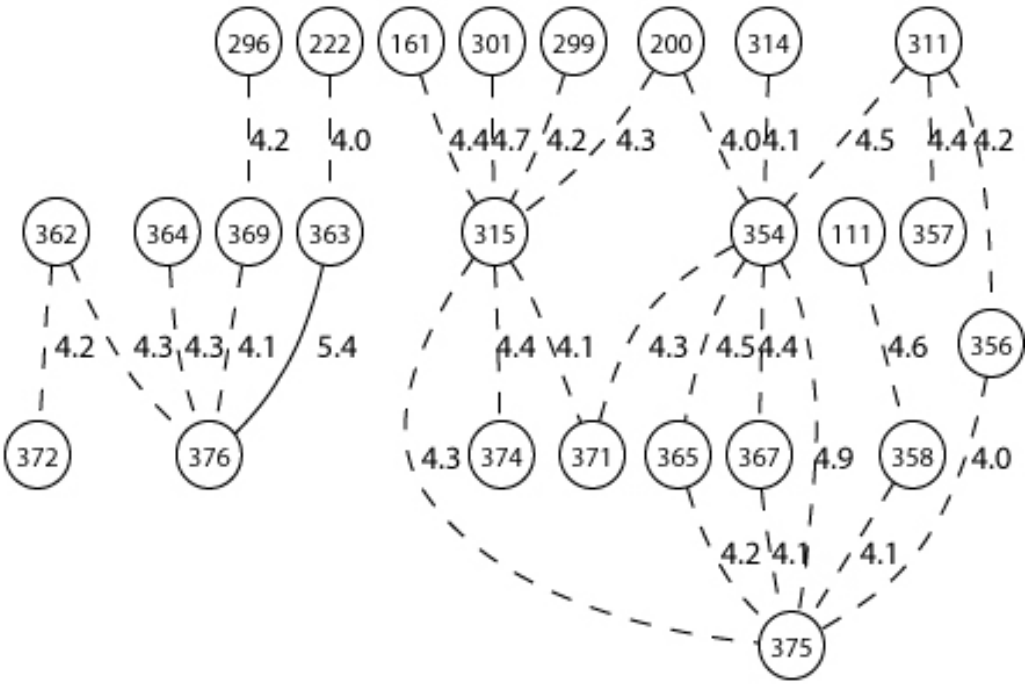

N - terminus
a
GTP_CDC
GTP_CDC
b
Septin Unique
C-terminus
c
Coiled-coil

Figure 4.2. The co-evolution network within each septin from septin Cdc3p, Cdc12p, Cdc10p, Cdc11p and Shs1p. One circle in a graph represents one amino acid. The number within the circle is the position of the residue in the corresponding septin sequence in *S. cerevisiae.* The values on the connecting lines are the Z-score values for mutual information. The thicker the connecting line, the higher the Z-score. (a) The co-evolution network of Cdc3p-Cdc3p (b) The co-evolution network of Cdc12p-Cdc12p (c) The co-evolution network of Cdc10p-Cdc10p (d) The co-evolution network of Cdc11p-Cdc11p (e) The co-evolution network of Shs1p-Shs1p


(a)
328
330
4.0
4.1
4.1
4.3
108
228
420
416
419
415
329
423
513
4.0 4.6 4.2
5.2 4.2 4.8 4.7 4.7
4.5 4.5 4.3 4.1 4.5 4.3 4.5
5.1 4.3 4.3
429
437
425
428
426
510
4.7
4.4
4.2 4.4 4.3 4.6 4.4
4.6
491
441
452
453
448
509
(b)
131
344
128
132
72
4.3 4.2 4.2 4.2 4.2 4.7
5.0 4.8 4.7 4.6 4.6 4.5 4.4 4.3 4.3 4.3 4.3
5.1
355
4.0
366
4.3 4.0 4.4
368
346
375
327
330
380
250
351
338
385
371
4.5
4.0 4.4
392
373
374
(c)
62
63
26
21
220
4.9 4.3 4.1 5.0
4.0
69
118
22
319
4.0 4.8 5.3 4.2 4.1 4.1 4.1 4.0 4.0 4.3 4.9 4.5 4.3
222
126
304
208
302
188
301
130
131
4.4
127
4.3 4.3
4.7
4.6
243
136

(c)


296
222
161
301
299
200
314
311
4.2
4.0
4.4
4.7
4.2
4.3
4.0
4.1
4.5
4.4
4.2
362
364
369
363
315
354
111
357
356
4.2
4.3
4.3
4.1
5.4
4.4
4.1
4.3
4.5
4.4
4.6
372
376
4.3
374
371
365
367
4.9
358
4.0
4.2
4.1
4.1
375

(d)

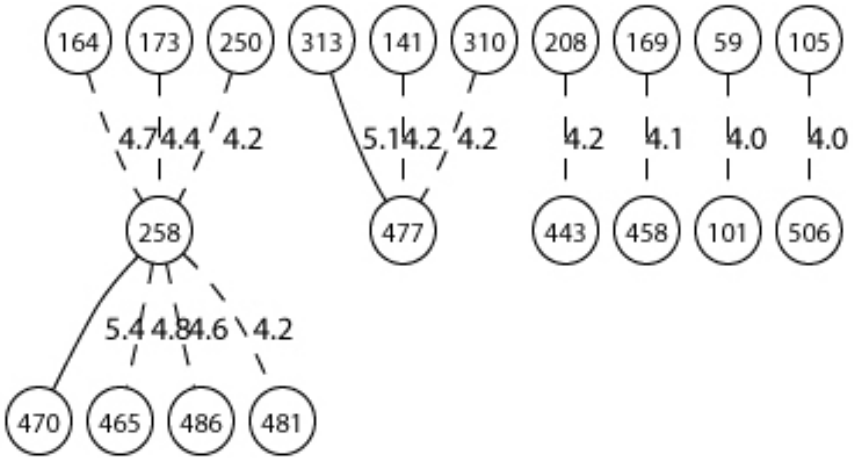

164
173
250
313
141
310
208
169
59
105
4.7
4.4
4.2
5.1
4.2
4.2
4.2
4.1
4.0
4.0
258
477
443
458
101
506
5.4
4.8
4.6
4.2
470
465
486
481

Figure 4.3. Predicted self-interactions within each septin protein sequence, Cdc3p, Cdc12p, Cdc10p, Cdc11p and Shs1p. The number within each circle is the residue position in each corresponding septin sequence in *S. cerevisiae.* The values on the connecting lines are the Z-score values of mutual information. The thicker the connecting line, the higher the Z-score. The letters represent the sequence domains that residues fall in. N: variable N-terminus. G is the GTP_CDC domain. U: the septin unique domain. VC: the variable C-terminus. C: the predicted coiled-coil region. (a) Residues from Cdc3p that are involved in Cdc3p self-interaction. Those amino acids are H329 Q416, S428, K429, and F452. (b) Cdc12p, residues are I72, E346 and E368. (c) Cdc10p, four amino acids are I22, L69, R127 and K304 (d) Cdc11p, three residues are E363, E369, F376 (e) Shs1p, the five residues are E258, N310Y465, Q477 and S486.

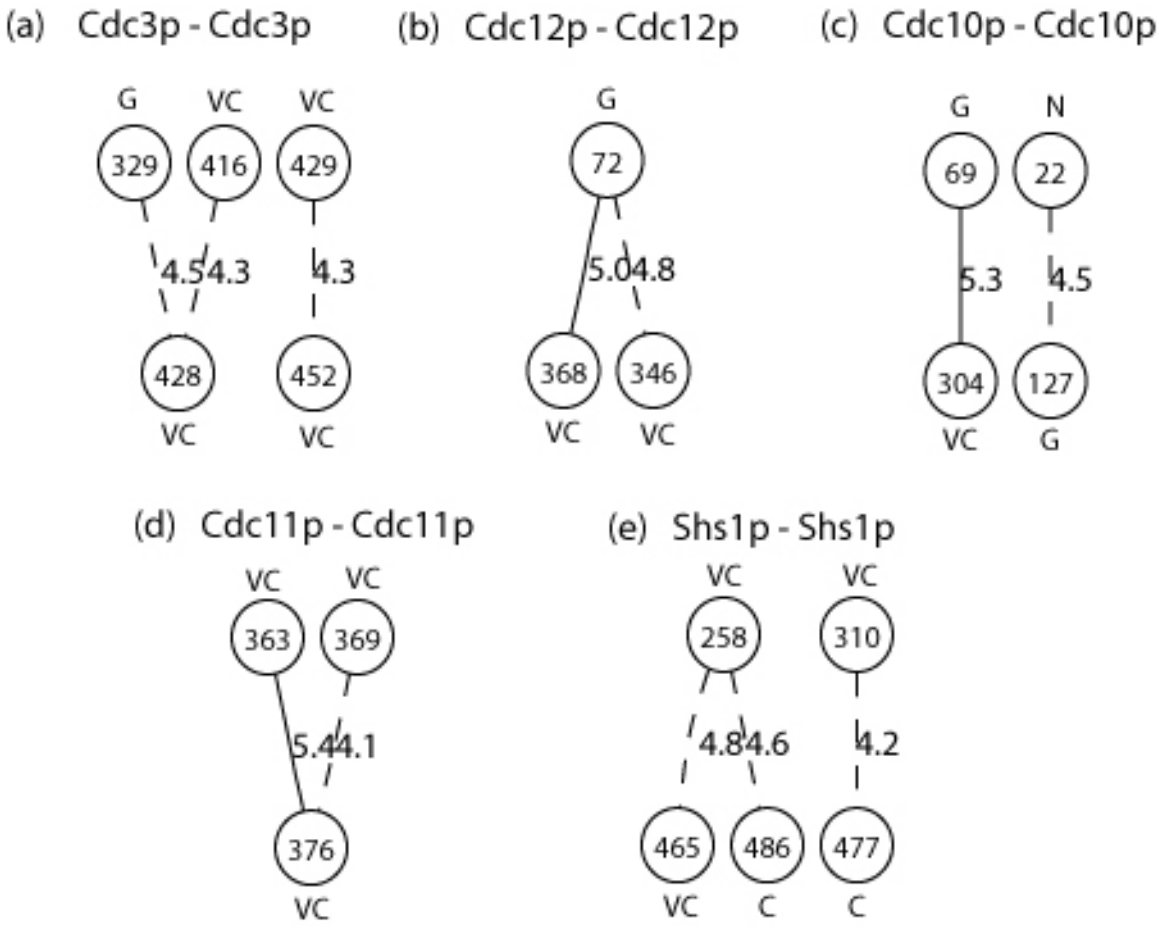

(a) Cdc3p - Cdc3p
G
VC
VC
329
416
429
4.5
4.3
4.3
428
452
VC
VC
(b) Cdc12p - Cdc12p
G
72
5.0
4.8
368
346
VC
VC
(c) Cdc10p - Cdc10p
G
N
69
22
5.3
4.5
304
127
VC
G
(d) Cdc11p - Cdc11p
VC
VC
363
369
5.4
4.1
376
VC
(e) Shs1p - Shs1p
VC
VC
258
310
4.8
4.6
4.2
465
486
477
VC
C
C

Figure 4.4. Predicted interactions of co-evolved residues between Cdc3p and other four septins in the complex. In each figure, upper row is Cdc3p. The numbers in the circles are the residue's positions in the corresponding sequences. The numbers on lines are the Z score for mutual information. The higher the Z score, the thicker the line is. The letters represent the sequence domains that residues fall in. N is the variable N-terminus. G is the GTP_CDC domain, U is the septin unique domain, VC is the variable C-terminus and C is the predicted coiled-coil region. (a) Interactions between Cdc3p and Cdc12p (b) Interactions between Cdc3p and Cdc10p (c) Interactions between Cdc3p and Cdc11p (d) Interactions between Cdc3p and Shs1p

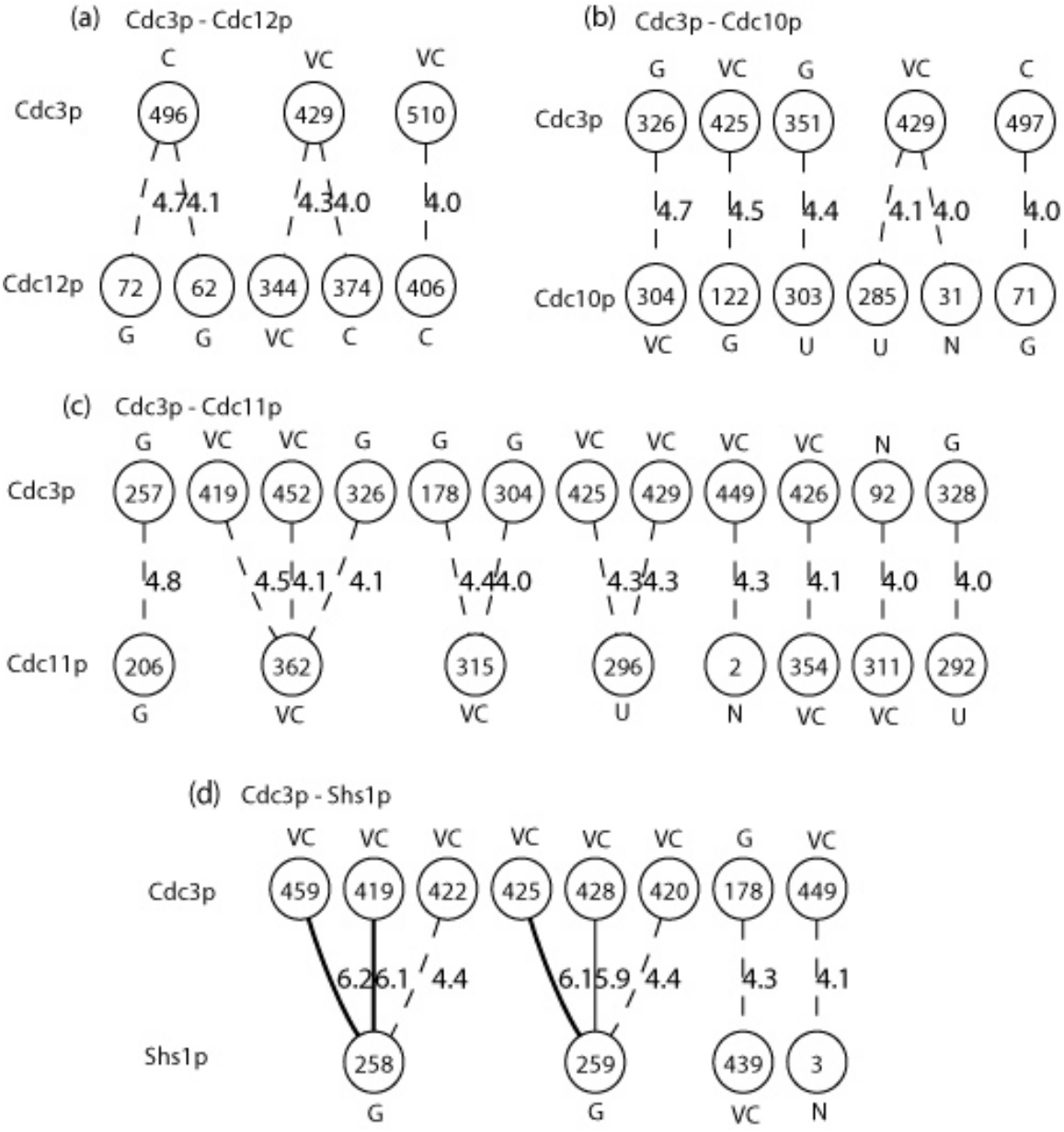

(a) Cdc3p - Cdc12p
C VC VC
Cdc3p
496 429 510
4.7 4.1 4.3 4.0 4.0
Cdc12p
72 62 344 374 406
G G VC C C
(b) Cdc3p - Cdc10p
G VC G VC C
Cdc3p
326 425 351 429 497
4.7 4.5 4.4 4.1 4.0 4.0
Cdc10p
304 122 303 285 31 71
VC G U U N G
(c) Cdc3p - Cdc11p
G VC VC G G G VC VC VC VC N G
Cdc3p
257 419 452 326 178 304 425 429 449 426 92 328
4.8 4.5 4.1 4.1 4.4 4.0 4.3 4.3 4.3 4.1 4.0 4.0
Cdc11p
206 362 315 296 2 354 311 292
G VC VC U N VC VC U
(d) Cdc3p - Shs1p
VC VC VC VC VC VC G VC
Cdc3p
459 419 422 425 428 420 178 449
6.2 6.1 4.4 6.1 5.9 4.4 4.3 4.1
Shs1p
258 259 439 3
G G VC N

Figure 4.5. Portion of alignment of Cdc3 vs. Cdc12 file. (a) Residue K429 (alignment position 649) is co-evolving with the residues E344 (alignment position 1839) and E374 (alignment position 11869) in sequence Cdc12. (b) These three columns are listed side by side. Column 1 is the K429 in Cdc3, column 2 is the E344 in Cdc12 and column 3 is the E 374 in Cdc12.

(a)


gi|Coccidioides_posadasii
gi|Mucor_circinelloides
gi|lactis_NRRL_Y-1140
|Ashbya_gossypii_ATCC_10895
gi|Candida_glabrata_CBS_138
gi|Saccharomyces_cerevisiae
gi|Candida_albicans
gi|Candida_albicans_SC5314
i|elongisporus_NRRL_YB-4239
gi|Pichia_stipitis_CBS_6054
Debaryomyces_hansenii_CBS76
gi|guilliermondii_ATCC_6260
Aspergillus_clavatus_NRRL_1
gi|Aspergillus_oryzae
Aspergillus_fumigatus_Af293
Neosartorya_fischeri_NRRL_1
Aspergillus_terreus_NIH2624
Aspergillus_nidulans_FGSC_A
gi|Emericella_nidulans
Aspergillus_niger_CBS_513.8
gi|Aspergillus_niger
gi|Coccidioides_immitis_RS
gi|Coccidioides_immitis
gi|Exophiala_dermatitidis
gi|Magnaporthe_grisea_70-15
gi|Magnaporthe_grisea
gi|Neurospora_crassa_OR74A
gi|Neurospora_crassa
Chaetomium_globosum_CBS_148
gi|Gibberella_zeae_PH-1
|Phaeosphaeria_nodorum_SN15
Cryptococcus_neoformans_JEC
gi|neoformans_B-3501A
gi|cinerea_okayama7#130
gi|Ustilago_maydis_521
Yarrowia_lipolytica_CLIB122
gi|pombe_972h-
i|Schizosaccharomyces_pombe
gi|Drosophila_pseudoobscura
gi|Drosophila_melanogaster
gi|Aedes_aegypti
Anopheles_gambiae_str._PEST
gi|Apis_mellifera
gi|Tribolium_castaneum
gi|Rattus_norvegicus
gi|Mus_musculus
gi|Macaca_fascicularis
gi|Canis_lupus_familiaris
gi|Gallus_gallus
gi|Bos_taurus
gi|Pongo_pygmaeus
gi|Homo_sapiens
gi|Macaca_mulatta
gi|Pan_troglodytes
gi|Pan_troglodytes_verus
gi|Xenopus_laevis
gi|Monodelphis_domestica
gi|Danio_rerio
Strongylocentrotus_purpurat
gi|Bombyx_mori
gi|synthetic_construct
gi|Geodia_cydonium
gi|Ornithorhynchus_anatinus
gi|Tetraodon_nigroviridis
gi|Xenopus_tropicalis
gi|Schistosoma_japonicum
gi|Caenorhabditis_elegans
gi|Caenorhabditis_briggsae
ruler
....640.....650...
....1830.....1840.....1850.....1860.....1870..
CDC 3
CDC 12

(b)

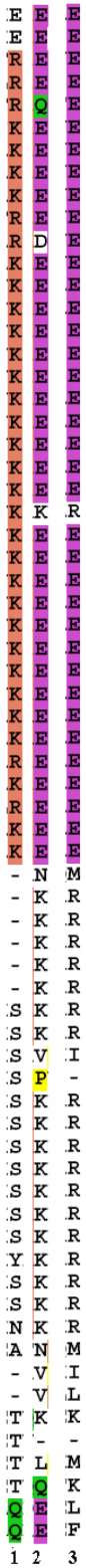
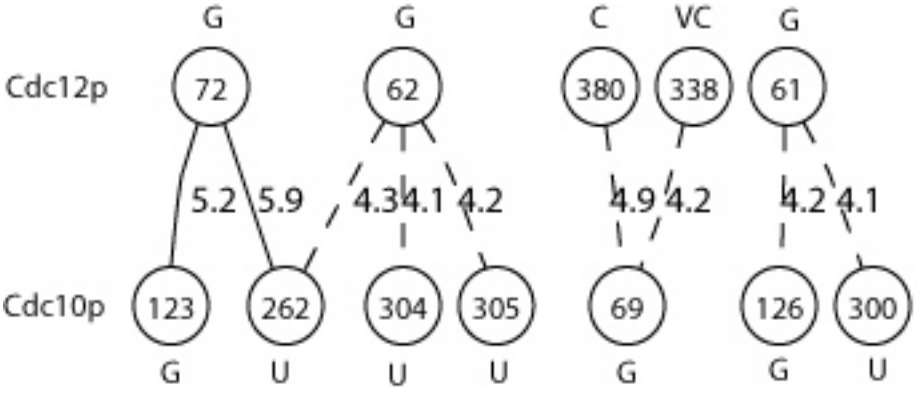
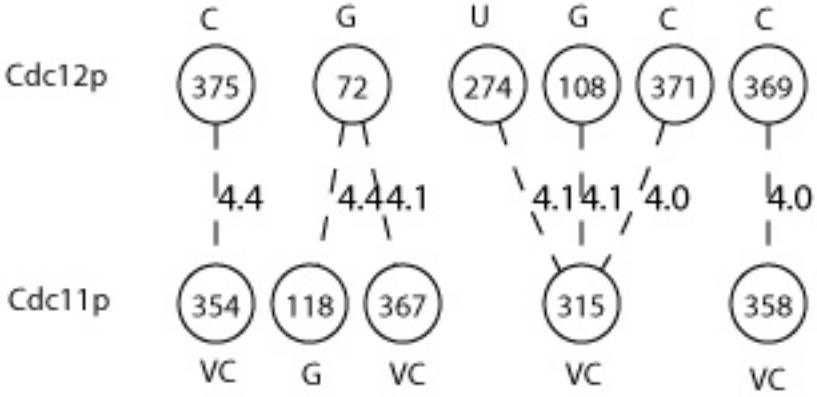

E E E
E E E
R E E
R E E
R Q E
K E E
K E E
K E E
K E E
R E E
R D E
K E E
K E E
K E E
K E E
K E E
K E E
K E E
K E E
K E E
K K R
K E E
K E E
K E E
K E E
K E E
K E E
K E E
K E E
K E E
K E E
R E E
K E E
R E E
K E E
- N M
- K R
- K R
- K R
- K R
- K R
S K R
S K R
S V I
S P -
S K R
S K R
S K R
S K R
S K R
S K R
S K R
Y K R
S K R
S K R
N K R
A N M
- V I
- V L
T K K
T - -
T L M
T Q K
Q E L
Q E F
1 2 3

Figure 4.6. . Predicted interactions of co-evolved residues between Cdc12p and other three septins in the complex. In each figure, upper row is Cdc12p. The numbers in the circles , the numbers on lines and the letters represent the same as in figure 4. (a) Interactions between Cdc12p and Cdc10p (b) Interactions between Cdc12p and Cdc11p (c) Interactions between Cdc12p and Shs1p

(a) Cdc12p - Cdc10p



(b) Cdc12p -- Cdc11p



(c) Cdc12p - Shs1p

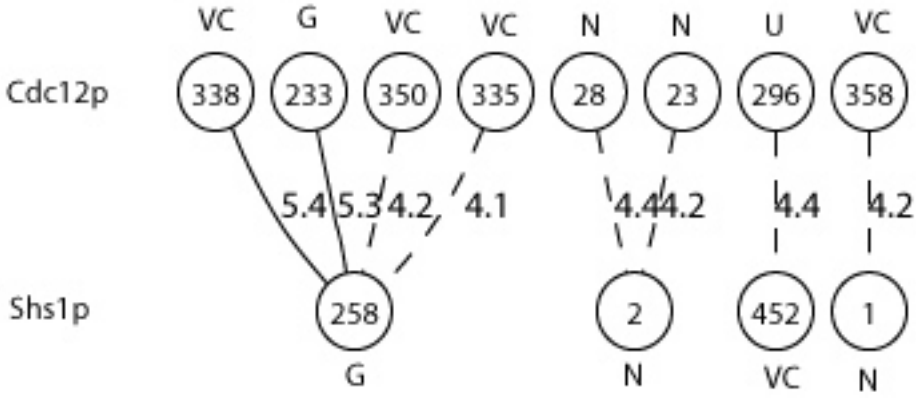


Figure 4.7. Predicted interactions of co-evolved residues between Cdc10p and other two septins in the complex. In each figure, upper row is Cdc10p. The numbers in the circles, the numbers on lines and the letters, represent the same as shown in figure 4. (a) Interactions between Cdc10p and Cdc11p (b) Interactions between Cdc10p and Shs1p

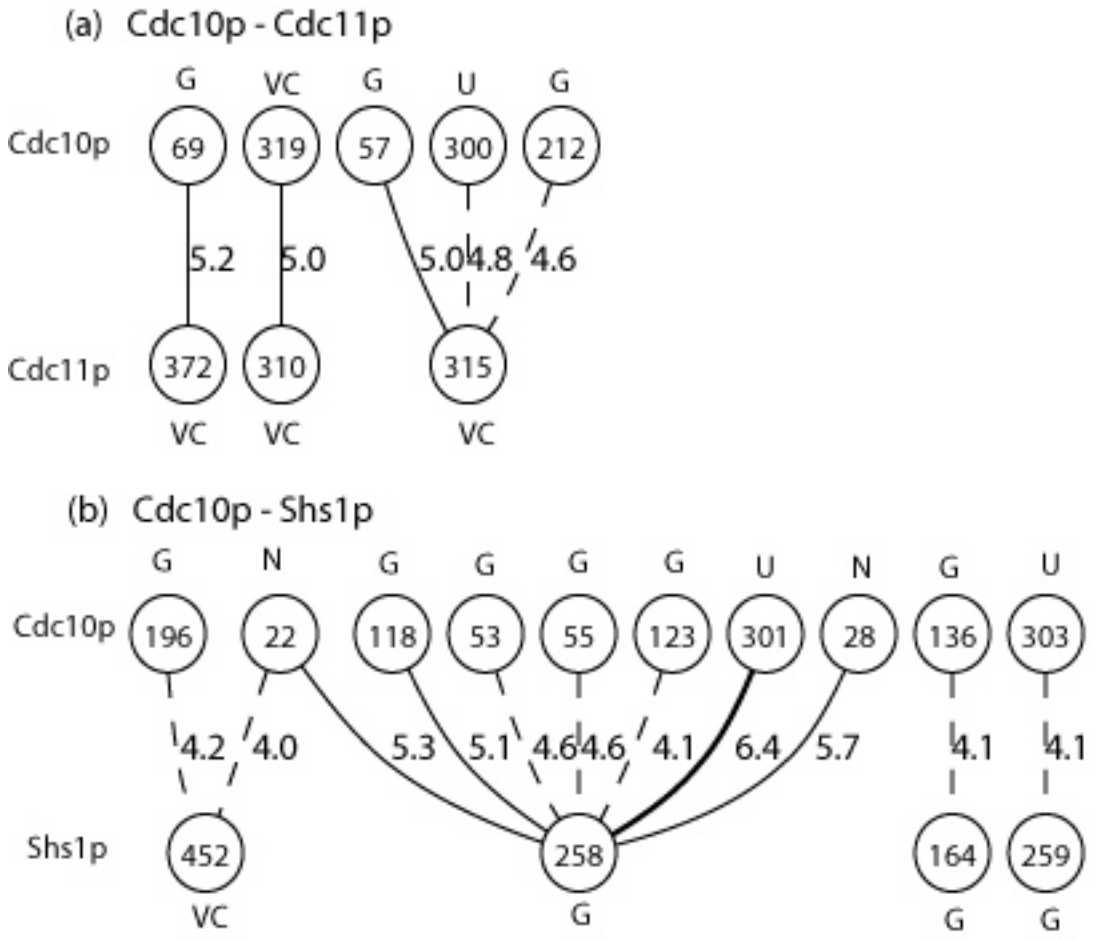

(a) Cdc10p - Cdc11p
G VC G U G
Cdc10p
69 319 57 300 212
5.2 5.0 5.0 4.8 4.6
Cdc11p
372 310 315
VC VC VC
(b) Cdc10p - Shs1p
G N G G G G U N G U
Cdc10p
196 22 118 53 55 123 301 28 136 303
4.2 4.0 5.3 5.1 4.6 4.6 4.1 6.4 5.7 4.1 4.1
Shs1p
452 258 164 259
VC G G G

Figure 4.8. Predicted interactions of co-evolved residues between Cdc11p and Shs1p. Top row is Cdc11p and bottom row is Shs1p. The numbers in the circles, the numbers on lines and the letters, represent the same as shown in figure 4.

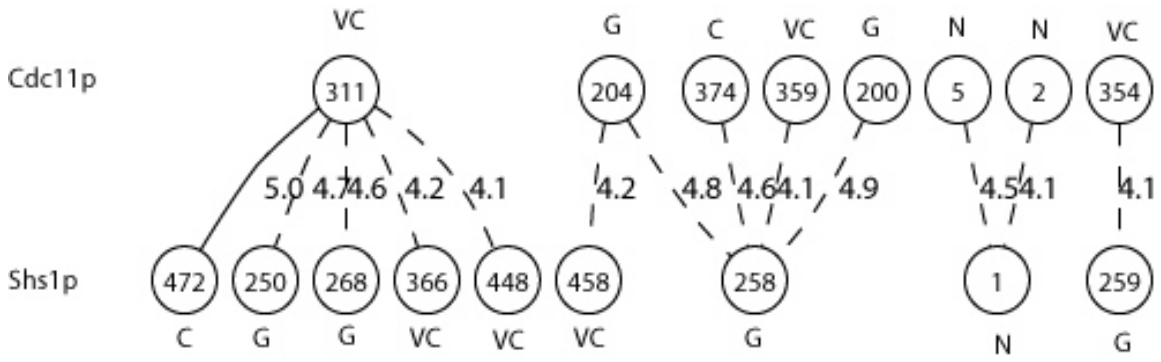

Cdc11p
Shs1p
VC
311
G
204
C
374
VC
359
G
200
N
5
N
2
VC
354
5.0
4.7
4.6
4.2
4.1
4.2
4.8
4.6
4.1
4.9
4.5
4.1
4.1
472
250
268
366
448
458
258
1
259
C
G
G
VC
VC
VC
G
N
G

Figure 4.9. Predicted interactions between formin (Bni1p) and Cdc3p, Cdc10p, Cdc11p and Shs1p.The number in each circle is the residue position and number on each line is the Z score for mutual information. Thicker line represents larger z score. In each sequence, Bni1p is the sequence at the bottom line. (a)Cdc3p with Bni1p (b) Cdc10p with Bni1p (c) Cdc11p with Bni1p

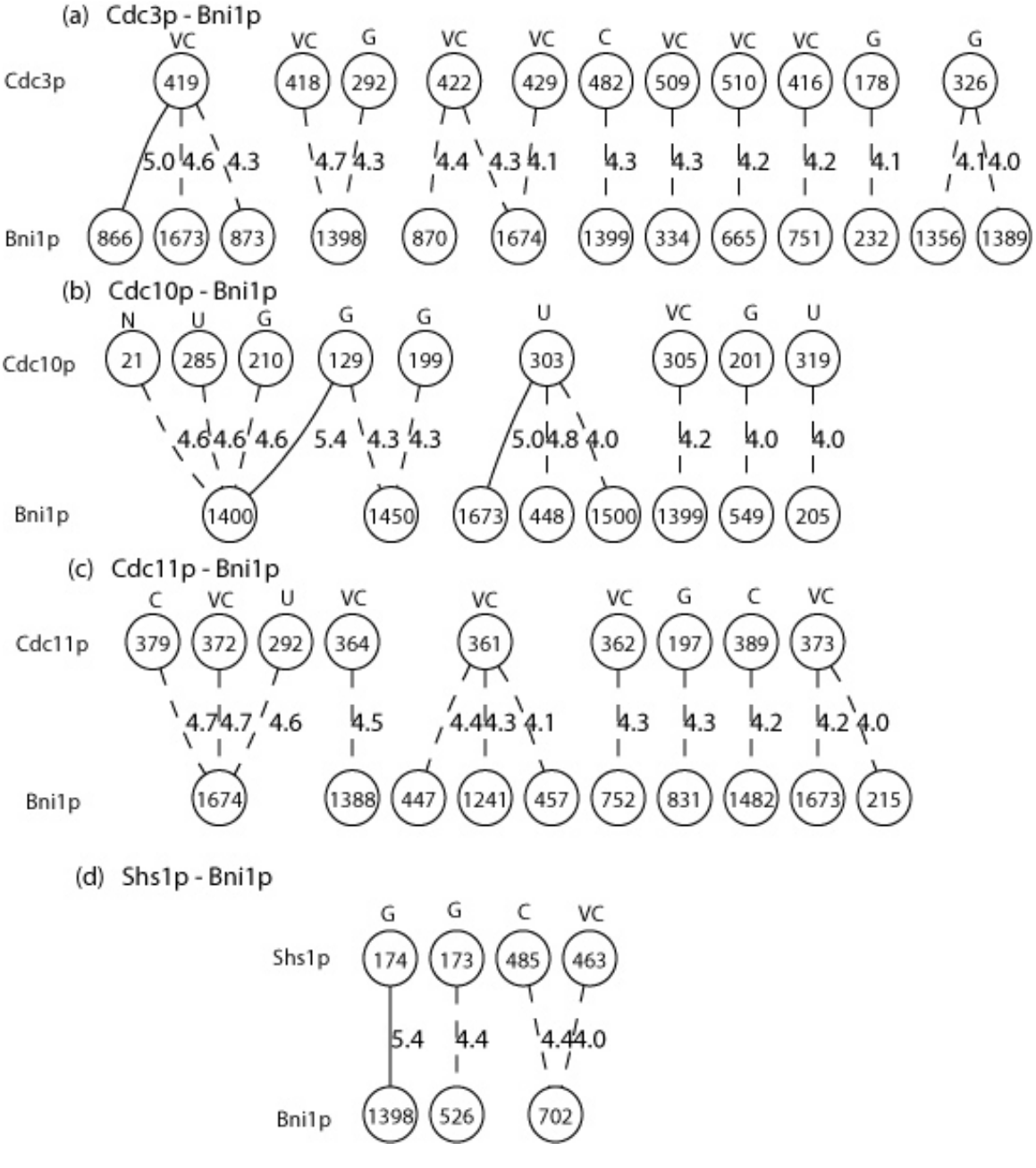

(a) Cdc3p - Bni1p
Cdc3p
VC 419
VC 418
G 292
VC 422
VC 429
C 482
VC 509
VC 510
VC 416
G 178
G 326
5.0 4.6 4.3
4.7 4.3
4.4 4.3 4.1
4.3
4.3
4.2
4.2
4.1
4.1 4.0
Bni1p
866 1673 873 1398 870 1674 1399 334 665 751 232 1356 1389
(b) Cdc10p - Bni1p
Cdc10p
N 21
U 285
G 210
G 129
G 199
U 303
VC 305
G 201
U 319
4.6 4.6 4.6 5.4 4.3 4.3
5.0 4.8 4.0
4.2
4.0
4.0
Bni1p
1400 1450 1673 448 1500 1399 549 205
(c) Cdc11p - Bni1p
Cdc11p
C 379
VC 372
U 292
VC 364
VC 361
VC 362
G 197
C 389
VC 373
4.7 4.7 4.6
4.5
4.4 4.3 4.1
4.3
4.3
4.2
4.2 4.0
Bni1p
1674 1388 447 1241 457 752 831 1482 1673 215
(d) Shs1p - Bni1p
Shs1p
G 174
G 173
C 485
VC 463
5.4
4.4
4.4 4.0
Bni1p
1398 526 702

Figure 4.10. Predicted interactions of co-evolving residues between Cdc12p and formin protein. The upper row is Cdc12p and the bottom row is formin Bni1p. The interactions form one large network and a separate small network. The number in each circle is the residue position and number on each line is the Z score for mutual information. Thicker line represents larger z score.

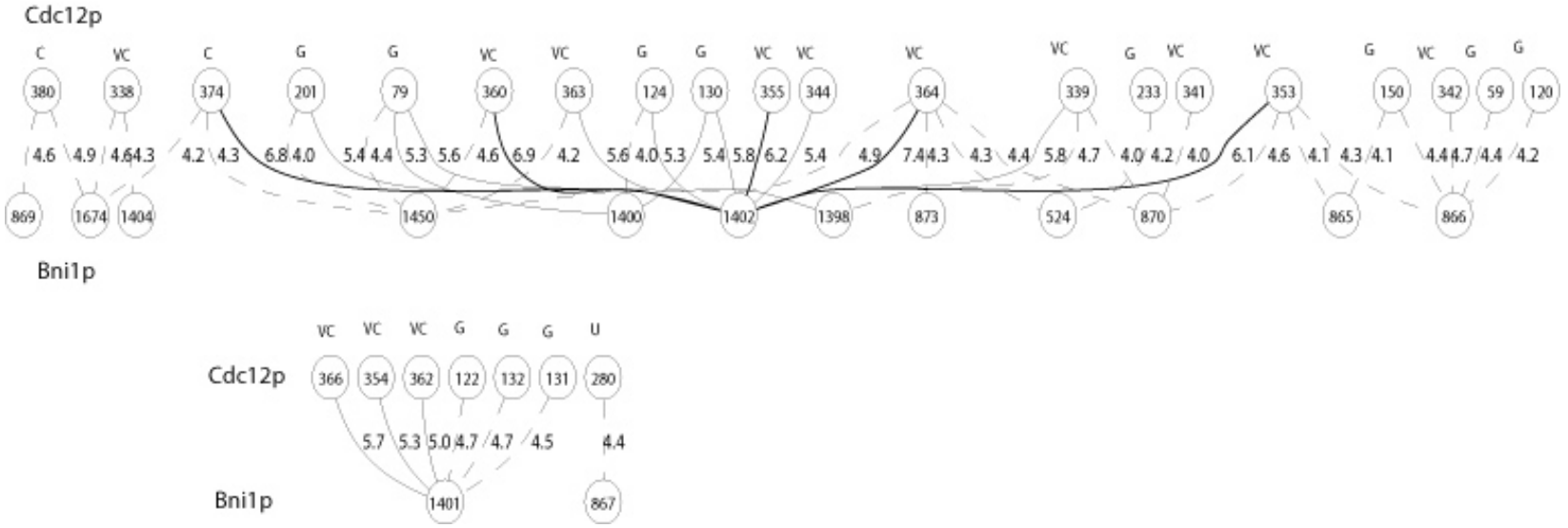

Cdc12p
C VC C G G VC VC G G VC VC VC VC G VC G G
380 338 374 201 79 360 363 124 130 355 344 364 339 233 341 353 150 342 59 120
4.6 4.9 4.6 4.3 4.2 4.3 6.8 4.0 5.4 4.4 5.3 5.6 4.6 6.9 4.2 5.6 4.0 5.3 5.4 5.8 6.2 5.4 4.9 7.4 4.3 4.3 4.4 5.8 4.7 4.0 4.2 4.0 6.1 4.6 4.1 4.3 4.1 4.4 4.7 4.4 4.2
869 1674 1404 1450 1400 1402 1398 873 524 870 865 866
Bni1p
VC VC VC G G G U
Cdc12p 366 354 362 122 132 131 280
5.7 5.3 5.0 4.7 4.7 4.5 4.4
Bni1p 1401 867

Figure 4.11. The interactions between five septin proteins and myosin. The symbols are the same as in previous figures. In each figure, myosin is the sequence at the bottom. (a) Cdc3p with Myo1p (b)Cdc12p with Myo1p (c) Cdc10p with Myo1p (d) Cdc11p with Myo1p (e) Shs1p with Myo1p

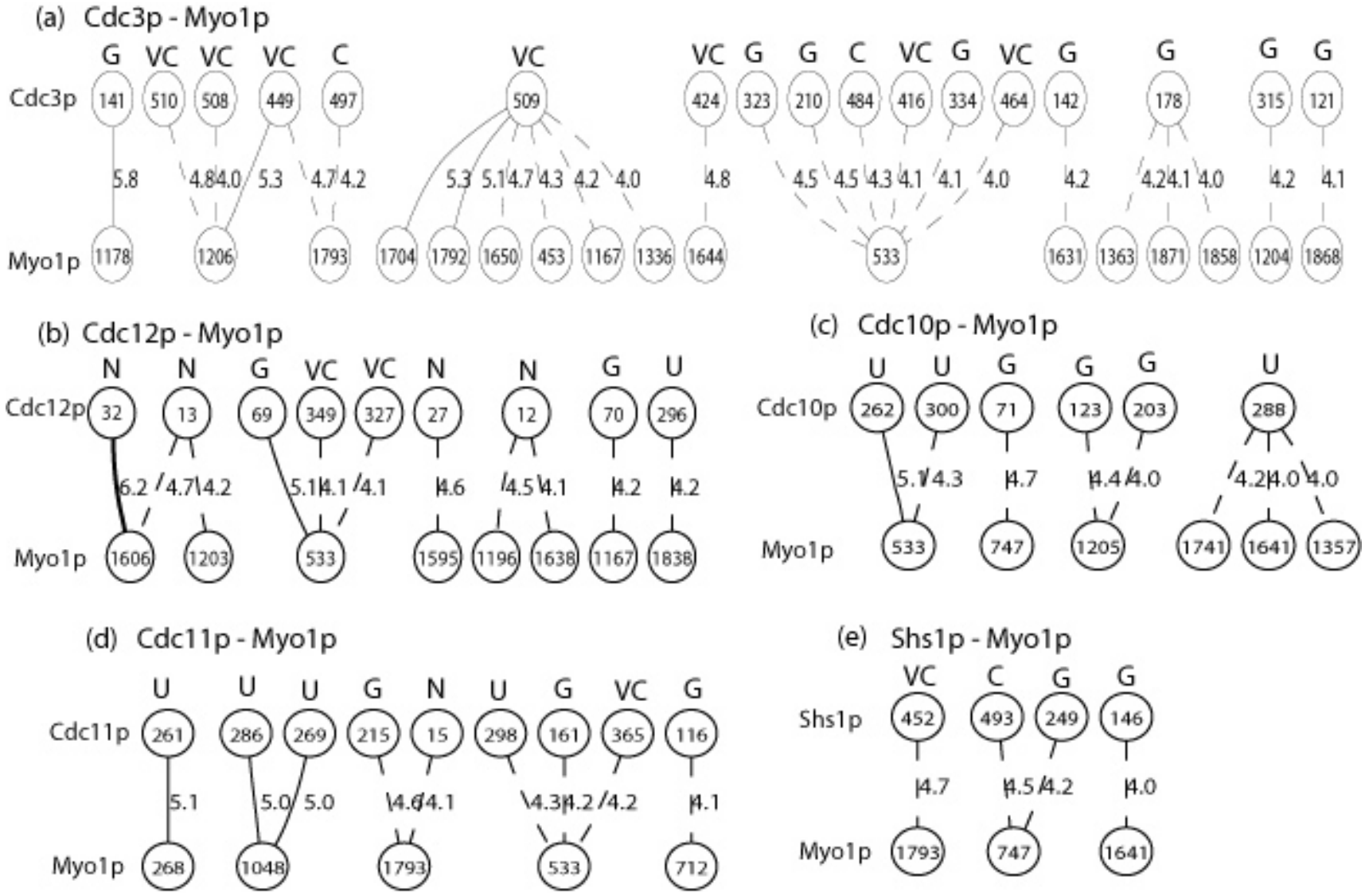

(a) Cdc3p - Myo1p
Cdc3p
Myo1p
(b) Cdc12p - Myo1p
Cdc12p
Myo1p
(c) Cdc10p - Myo1p
Cdc10p
Myo1p
(d) Cdc11p - Myo1p
Cdc11p
Myo1p
(e) Shs1p - Myo1p
Shs1p
Myo1p

Table 4.1. Number of predicted co-evolving pairs within or between two proteins

	Cdc3p	**Cdc12p**	**Cdc10p**	**Cdc11p**	**Shs1p**	**Bni1p**	**Myo1p**
Cdc3p	30 (0.01)	41 (0.01)	82 (0.02)	76 (0.02)	58 (0.02)	114 (0.03)	220 (0.07)
Cdc12p	41 (0.01)	25 (0.01)	68 (0.02)	59 (0.02)	54 (0.02)	252 (0.07)	114 (0.03)
Cdc10p	82 (0.02)	68 (0.02)	23 (0.01)	53 (0.02)	51 (0.02)	65 (0.02)	45 (0.01)
Cdc11p	76 (0.02)	59 (0.02)	53 (0.02)	28 (0.01)	83 (0.02)	61 (0.02)	46 (0.01)
Shs1p	58 (0.02)	54 (0.02)	51 (0.02)	83 (0.02)	14 (0)	42 (0.01)	40 (0.01)
Bni1p	114 (0.03)	252 (0.07)	65 (0.02)	61 (0.02)	42 (0.01)	not tested	not tested
Myo1p	220 (0.07)	114 (0.03)	45 (0.01)	46 (0.01)	40 (0.01)	not tested	not tested

The number in each cell is the predicted co-evolving pairs based on mutual information. The number in paranthesis is the percentage out of the sum of all the numbers in table. There are 30 co-evolving pairs within Cdc3p protein, and 220 co-evolving pairs between Cdc3p and Myo1p.

Table 4.2. Number of predicted interaction pairs within or between two proteins.

	Cdc3p	**Cdc12p**	**Cdc10p**	**Cdc11p**	**Shs1p**	**Bni1p**	**Myo1p**
Cdc3p	3 (0.01)	5 (0.01)	6 (0.01)	12 (0.02)	8 (0.02)	15 (0.03)	25 (0.05)
Cdc12p	5 (0.01)	2 (0)	9 (0.02)	7 (0.01)	8 (0.02)	48 (0.09)	11 (0.02)
Cdc10p	6 (0.01)	9 (0.02)	2 (0)	5 (0.01)	11 (0.02)	12 (0.02)	18 (0.03)
Cdc11p	12 (0.02)	7 (0.01)	5 (0.01)	2 (0)	13 (0.03)	12 (0.02)	19 (0.04)
Shs1p	8 (0.02)	8 (0.02)	11 (0.02)	13 (0.03)	3 (0.01)	4 (0.01)	4 (0.01)
Bni1p	15 (0.03)	48 (0.09)	12 (0.02)	12 (0.02)	4 (0.01)	not tested	not tested
Myo1p	25 (0.05)	11 (0.02)	18 (0.03)	19 (0.04)	4 (0.01)	not tested	not tested

The number in each cell is the predicted co-evolving pairs that are in physical contact in 3D structure among those with high mutual information. The number in paranthesis is the percentage out of the sum of all the numbers in table. There are 3 interacting co-evolving pairs within Cdc3p protein, and 25 interacting co-evolving pairs between Cdc3p and Myo1p.

CHAPTER 5

CONCLUSIONS

The phylogeny of septins shows orthology across kingdoms

Separate evolutionary histories for the septin gene families from animal and fungal kingdoms were suggested previously [115]. Under this hypothesis, all fungal sequences would be more closely related to each other than to any animal sequences, and vice versa. With more and more septin sequences coming out, especially in a wider variety of species, information was available for a more thorough evolutionary analysis of the septin family. In Chapter 2, we presented a phylogenetic analysis of the septin gene family based on 162 septins from fungi, microsporidia and animals. We: 1) reconstructed the evolutionary history of septins and showed the existence of orthologous relationships between animal and fungal septins; 2) categorized all septins into five big groups, GROUPS 1~5 of which two of the groups contained septins from animals and fungi, two other groups contained septins only from fungi and one group contained septins from fungi and microsporidia; 3) suggested consistent nomenclature for septins according to the groups to which they belonged**.** This phylogeny should help future experimental and comparative septin analysis across kingdoms. As septins are most well studied in *Saccharomyces cerevisiae*, findings of septin functions in yeast may indicate similar functions for corresponding septin orthologs in animals. However, the origin of the septin family cannot yet be clearly resolved from the information available. Septin sequences from organisms at the base of the evolutionary tree are needed.

Conserved motifs and positions in septins

Septins belong to the GTPase super class [116]. Though certain amino acid sequences and motifs are conserved, the overall identity varies a lot within this family. In Chapter 2, we identified conserved positions in septin protein sequences based on a sequence alignment of 162 septins. We were able to extend the already known G1, G3, and G4 GTPase motifs within the GTP_CDC domain and find four conserved motifs and six conserved single positions in septins. Among the four newly discovered motifs, one was a septin specific motif.

Protein residue interaction prediction using mutual information

Proteins can work in a large interaction network and these interactions between proteins are often preserved across evolution. Conserved positions on protein sequences are likely to be important for protein functions and interactions. However, two proteins can both mutate such that the protein interactions are still kept. Thus, two interacting proteins may show a pattern of co-evolution across many species. In Chapter 3, we presented co-evolution analyses using mutual information statistics and applied this to protein contact prediction. We extracted sequences and 3D structure information for 48 pairs of interacting proteins from the Protein Complex Crystallization Database. We were able to: 1) show that co-evolving amino acid residue pairs were physically closer to each other than random pairs; 2) decide residue co-evolving pairing preference matrices and showed different amino acid preferences of co-evolution; 3) develop contact scoring matrices for co-evolving residues to computationally predict physical closeness between two co-evolving residues. As with most of the computational methods in this area, the sensitivity and specificity were not that high, but our approach improved on existing methods, and provided a new direction for residue interaction prediction based on co-evolution.

The application of mutual information analysis to the septin gene family

Some self-interactions were detected within a septin, but the numbers were much smaller than those of the interactions between different subunits from the septin filament complex. Septins have four major variable regions: the N-terminus; the variable region within the GTP_CDC domain; the variable region at the septin unique domain; and, the variable C-terminus. In Chapter 4, we concluded that the co-evolving interactions between two different septin units had preferences for different variable regions. The co-evolving interactions were more often detected in some variable regions than the others in septins.

In Chapter 4 we also suggested that two septin proteins were found to be very likely to interact with formin and myosin. Those two septins' numbers of co-evolving residues and co-evolutionarily interacting sites with formin and myosin were much larger than were those of any of the other septins. Septin Cdc12p, rather than the other component members of septin complex, was predicted to be more likely to interact with Bni1p, a member of formin family. This prediction result agreed with previous work showing that Bni1 interacted with septin Cdc12 in yeast [113]. Septin Cdc3p was predicted to be more likely to interact with Myo1p, a member of myosin family. Previous experiments showed that a mutation in Cdc12p abolished the localization of Myo1p, however no other information of direct interaction between Cdc12p and Myo1p was available [110]. We suggested that the Myo1p was interacting with the Cdc3p unit in the Cdc3-Cdc12-Cdc11-Cdc10-Shs1 septin complex.

Bioinformatics relies on the accuracy of data and can retrospectively help to improve data curation. Large scale genome sequencing projects have provided large amounts of sequence information. The accuracy of the data curation is important but is also a challenge. Wrong interpretations lead to wasted experiments and time. In septin phylogeny, our results depended

on database information from GenBank. However, we could have reconstructed a better evolutionary tree if sequence information for organisms near the base of evolutionary tree was available and accurate. On the other side, our analysis showed orthologous relationships between septins in different organisms and suggested consistent nomenclature, which could help to interpret future sequence data.

Implications for future work with septins and other protein families

The amino acid residue positions identified in Chapters 2 and 4 are likely to be very important for septins, either for their structures or protein interactions. Some future work applying our findings might include possible mutagenesis experiments on septin protein sequences targeting on the positions that show either conservation or co-evolving interactions.

The phylogenetic methods we used are well established, but our use of mutual information for prediction is relatively new. Better methods that can improve residue interaction prediction accuracy are certainly worth investigation. The pairing preference matrices for interactions showed some differences between interactions within one protein subunit and between two different protein subunits. As of now, however, there isn't enough data to generate separate matrices for interactions both between two subunits and within a subunit. In the future it might be possible to derive different prediction matrices for these different types of interactions when more data is available from the Protein Data Bank. This will increase the interaction prediction accuracy. The same method could also be applied to other protein families.

REFERENCES

1. Roos DS: **Computational biology. Bioinformatics--trying to swim in a sea of data**. *Science* 2001, **291**(5507):1260-1261.

2. Longtine MS, DeMarini DJ, Valencik ML, Al-Awar OS, Fares H, De Virgilio C, Pringle JR: **The septins: roles in cytokinesis and other processes**. *Curr Opin Cell Biol* 1996, **8**(1):106-119.

3. Pan F, Malmberg RL, Momany M: **Analysis of septins across kingdoms reveals orthology and new motifs**. *BMC evolutionary biology* 2007, **7**:103.

4. Versele M, Gullbrand B, Shulewitz MJ, Cid VJ, Bahmanyar S, Chen RE, Barth P, Alber T, Thorner J: **Protein-Protein Interactions Governing Septin Heteropentamer Assembly and Septin Filament Organization in Saccharomyces cerevisiae**. *Mol Biol Cell* 2004, **15**(10):4568-4583.

5. Kozubowski L, Larson JR, Tatchell K: **Role of the septin ring in the asymmetric localization of proteins at the mother-bud neck in Saccharomyces cerevisiae**. *Mol Biol Cell* 2005, **16**(8):3455-3466.

6. Fares H, Goetsch L, Pringle JR: **Identification of a developmentally regulated septin and involvement of the septins in spore formation in Saccharomyces cerevisiae**. *J Cell Biol* 1996, **132**(3):399-411.

7. Field CM, Kellogg D: **Septins: cytoskeletal polymers or signalling GTPases?** *Trends Cell Biol* 1999, **9**(10):387-394.

8. Lindsey R, Momany M: **Septin localization across kingdoms: three themes with variations**. *Curr Opin Microbiol* 2006, **9**(6):559-565.

9. Longtine MS, Fares H, Pringle JR: **Role of the Yeast Gin4p Protein Kinase in Septin Assembly and the Relationship between Septin Assembly and Septin Function**. *J Cell Biol* 1998, **143**(3):719-736.

10. Chant J, Pringle JR: **Patterns of bud-site selection in the yeast Saccharomyces cerevisiae**. *J Cell Biol* 1995, **129**(3):751-765.

11. Warenda AJ, Kauffman S, Sherrill TP, Becker JM, Konopka JB: **Candida albicans Septin Mutants Are Defective for Invasive Growth and Virulence**. *Infect Immun* 2003, **71**(7):4045-4051.

12. Giot L, Konopka JB: **Functional analysis of the interaction between Afr1p and the Cdc12p septin, two proteins involved in pheromone-induced morphogenesis**. *Mol Biol Cell* 1997, **8**(6):987-998.

13. Takizawa PA, DeRisi JL, Wilhelm JE, Vale RD: **Plasma Membrane Compartmentalization in Yeast by Messenger RNA Transport and a Septin Diffusion Barrier**. *Science* 2000, **290**(5490):341-344.

14. Sakchaisri K, Asano S, Yu L-R, Shulewitz MJ, Park CJ, Park J-E, Cho Y-W, Veenstra TD, Thorner J, Lee KS: **Coupling morphogenesis to mitotic entry**. *PNAS* 2004, **101**(12):4124-4129.

15. Kinoshita M, Kumar S, Mizoguchi A, Ide C, Kinoshita A, Haraguchi T, Hiraoka Y, Noda M: **Nedd5, a mammalian septin, is a novel cytoskeletal component interacting with actin-based structures**. *Genes Dev* 1997, **11**(12):1535-1547.

16. Surka MC, Tsang CW, Trimble WS: **The Mammalian Septin MSF Localizes with Microtubules and Is Required for Completion of Cytokinesis**. *Mol Biol Cell* 2002, **13**(10):3532-3545.

17. Hsu SC, Hazuka CD, Roth R, Foletti DL, Heuser J, Scheller RH: **Subunit composition, protein interactions, and structures of the mammalian brain sec6/8 complex and septin filaments**. *Neuron* 1998, **20**(6):1111-1122.

18. Martinez C, Ware J: **Mammalian Septin Function in Hemostasis and Beyond**. *Experimental Biology and Medicine* 2004, **229**(11):1111-1119.

19. Leipe DD, Wolf YI, Koonin EV, Aravind L: **Classification and evolution of P-loop GTPases and related ATPases**. *J Mol Biol* 2002, **317**(1):41-72.

20. Casamayor A, Snyder M: **Molecular dissection of a yeast septin: distinct domains are required for septin interaction, localization, and function**. *Mol Cell Biol* 2003, **23**(8):2762-2777.

21. Huelsenbeck JP, Ronquist F: **MRBAYES: Bayesian inference of phylogenetic trees**. *Bioinformatics (Oxford, England)* 2001, **17**(8):754-755.

22. Mau B, Newton MA, Larget B: **Bayesian phylogenetic inference via Markov chain Monte Carlo methods**. *Biometrics* 1999, **55**(1):1-12.

23. Huelsenbeck JP, Ronquist F, Nielsen R, Bollback JP: **Bayesian inference of phylogeny and its impact on evolutionary biology**. *Science (New York, NY* 2001, **294**(5550):2310-2314.

24. Crooks GE, Hon G, Chandonia JM, Brenner SE: **WebLogo: a sequence logo generator**. *Genome research* 2004, **14**(6):1188-1190.

25. Dimmic MW, Hubisz MJ, Bustamante CD, Nielsen R: **Detecting coevolving amino acid sites using Bayesian mutational mapping**. *Bioinformatics (Oxford, England)* 2005, **21 Suppl 1**:i126-135.

26. Dutheil J, Pupko T, Jean-Marie A, Galtier N: **A model-based approach for detecting coevolving positions in a molecule**. *Molecular biology and evolution* 2005, **22**(9):1919-1928.

27. Goh CS, Bogan AA, Joachimiak M, Walther D, Cohen FE: **Co-evolution of proteins with their interaction partners**. *Journal of molecular biology* 2000, **299**(2):283-293.

28. Goh CS, Cohen FE: **Co-evolutionary analysis reveals insights into protein-protein interactions**. *Journal of molecular biology* 2002, **324**(1):177-192.

29. Hamilton N, Burrage K, Ragan MA, Huber T: **Protein contact prediction using patterns of correlation**. *Proteins* 2004, **56**(4):679-684.

30. Gloor GB, Martin LC, Wahl LM, Dunn SD: **Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions**. *Biochemistry* 2005, **44**(19):7156-7165.

31. Fares MA, Travers SA: **A novel method for detecting intramolecular coevolution: adding a further dimension to selective constraints analyses**. *Genetics* 2006, **173**(1):9-23.

32. Fukami-Kobayashi K, Schreiber DR, Benner SA: **Detecting compensatory covariation signals in protein evolution using reconstructed ancestral sequences**. *Journal of molecular biology* 2002, **319**(3):729-743.

33. Oliveira L, Paiva AC, Vriend G: **Correlated mutation analyses on very large sequence families**. *Chembiochem* 2002, **3**(10):1010-1017.

34. Pazos F, Helmer-Citterich M, Ausiello G, Valencia A: **Correlated mutations contain information about protein-protein interaction**. *J Mol Biol* 1997, **271**(4):511-523.

35. Pritchard L, Bladon P, J MOM, M JD: **Evaluation of a novel method for the identification of coevolving protein residues**. *Protein engineering* 2001, **14**(8):549-555.

36. Tuff P, Darlu P: **Exploring a phylogenetic approach for the detection of correlated substitutions in proteins**. *Mol Biol Evol* 2000, **17**(11):1753-1759.

37. Tramontano A: **Protein structure prediction : concepts and applications**: Weinheim : Wiley-VCH ; Chichester : John Wiley [distributor]; 2006.

38. Fariselli P, Olmea O, Valencia A, Casadio R: **Progress in predicting inter-residue contacts of proteins with neural networks and correlated mutations**. *Proteins* 2001, **Suppl 5**:157-162.

39. Glaser F, Steinberg DM, Vakser IA, Ben-Tal N: **Residue frequencies and pairing preferences at protein-protein interfaces**. *Proteins* 2001, **43**(2):89-102.

40. Vicatos S, Reddy BV, Kaznessis Y: **Prediction of distant residue contacts with the use of evolutionary information**. *Proteins* 2005, **58**(4):935-949.

41. Singer MS, Vriend G, Bywater RP: **Prediction of protein residue contacts with a PDB-derived likelihood matrix**. *Protein Eng* 2002, **15**(9):721-725.

42. Song J, Burrage K: **Predicting residue-wise contact orders in proteins by support vector regression**. *BMC Bioinformatics* 2006, **7**(1):425.

43. Harris SD: **Septum formation in Aspergillus nidulans**. *Current opinion in microbiology* 2001, **4**(6):736-739.

44. Fares H, Goetsch L, Pringle JR: **Identification of a developmentally regulated septin and involvement of the septins in spore formation in Saccharomyces cerevisiae**. *J Cell Biol* 1996, **132**(3):399-411.

45. De Virgilio C, DeMarini DJ, Pringle JR: **SPR28, a sixth member of the septin gene family in Saccharomyces cerevisiae that is expressed specifically in sporulating cells**. *Microbiology* 1996, **142 (Pt 10)**:2897-2905.

46. Longtine MS, Bi E: **Regulation of septin organization and function in yeast**. *Trends Cell Biol* 2003, **13**(8):403-409.

47. Douglas LM, Alvarez FJ, McCreary C, Konopka JB: **Septin Function in Yeast Model Systems and Pathogenic Fungi**. *Eukaryotic Cell* 2005, **4**(9):1503-1512.

48. Martinez C, Ware J: **Mammalian septin function in hemostasis and beyond**. *Exp Biol Med (Maywood)* 2004, **229**(11):1111-1119.

49. Hall PA, Russell SE: **The pathobiology of the septin gene family**. *J Pathol* 2004, **204**(4):489-505.

50. Spiliotis ET, Kinoshita M, Nelson WJ: **A mitotic septin scaffold required for Mammalian chromosome congression and segregation**. *Science* 2005, **307**(5716):1781-1785.

51. Kinoshita M: **Diversity of septin scaffolds**. *Curr Opin Cell Biol* 2006, **18**(1):54-60.

52. Bourne HR, Sanders DA, McCormick F: **The GTPase superfamily: conserved structure and molecular mechanism**. *Nature* 1991, **349**(6305):117-127.

53. Saraste M, Sibbald PR, Wittinghofer A: **The P-loop--a common motif in ATP- and GTP-binding proteins**. *Trends Biochem Sci* 1990, **15**(11):430-434.

54. Vetter IR, Wittinghofer A: **Nucleoside triphosphate-binding proteins: different scaffolds to achieve phosphoryl transfer**. *Q Rev Biophys* 1999, **32**(1):1-56.

55. Dever TE, Glynias MJ, Merrick WC: **GTP-binding domain: three consensus sequence elements with distinct spacing**. *Proceedings of the National Academy of Sciences of the United States of America* 1987, **84**(7):1814-1818.

56. Field CM, al-Awar O, Rosenblatt J, Wong ML, Alberts B, Mitchison TJ: **A purified Drosophila septin complex forms filaments and exhibits GTPase activity**. *J Cell Biol* 1996, **133**(3):605-616.

57. Mendoza M, Hyman AA, Glotzer M: **GTP binding induces filament assembly of a recombinant septin**. *Curr Biol* 2002, **12**(21):1858-1863.

58. Vrabioiu AM, Gerber SA, Gygi SP, Field CM, Mitchison TJ: **The majority of the Saccharomyces cerevisiae septin complexes do not exchange guanine nucleotides**. *J Biol Chem* 2004, **279**(4):3111-3118.

59. Zhang J, Kong C, Xie H, McPherson PS, Grinstein S, Trimble WS: **Phosphatidylinositol polyphosphate binding to the mammalian septin H5 is modulated by GTP**. *Curr Biol* 1999, **9**(24):1458-1467.

60. Versele M, Gullbrand B, Shulewitz MJ, Cid VJ, Bahmanyar S, Chen RE, Barth P, Alber T, Thorner J: **Protein-protein interactions governing septin heteropentamer assembly and septin filament organization in Saccharomyces cerevisiae**. *Mol Biol Cell* 2004, **15**(10):4568-4583.

61. An H, Morrell JL, Jennings JL, Link AJ, Gould KL: **Requirements of fission yeast septins for complex formation, localization, and function**. *Mol Biol Cell* 2004, **15**(12):5551-5564.

62. Versele M, Thorner J: **Some assembly required: yeast septins provide the instruction manual**. *Trends in Cell Biology* 2005, **15**(8):414-424.

63. Momany M, Zhao J, Lindsey R, Westfall PJ: **Characterization of the Aspergillus nidulans septin (asp) gene family**. *Genetics* 2001, **157**(3):969-977.

64. Kinoshita M: **The septins**. *Genome Biol* 2003, **4**(11):236.

65. Schultz J, Milpetz F, Bork P, Ponting CP: **SMART, a simple modular architecture research tool: identification of signaling domains**. *Proceedings of the National Academy of Sciences of the United States of America* 1998, **95**(11):5857-5864.

66. Boyce KJ, Chang H, D'Souza CA, Kronstad JW: **An Ustilago maydis septin is required for filamentous growth in culture and for full symptom development on maize**. *Eukaryot Cell* 2005, **4**(12):2044-2056.

67. Hall PA, Jung K, Hillan KJ, Russell SE: **Expression profiling the human septin gene family**. *J Pathol* 2005, **206**(3):269-278.

68. Ronquist F, Huelsenbeck JP: **MrBayes 3: Bayesian phylogenetic inference under mixed models**. *Bioinformatics* 2003, **19**(12):1572-1574.

69. Kartmann B, Roth D: **Novel roles for mammalian septins: from vesicle trafficking to oncogenesis**. *J Cell Sci* 2001, **114**(Pt 5):839-844.

70. Kinoshita M: **Assembly of mammalian septins**. *J Biochem (Tokyo)* 2003, **134**(4):491-496.

71. Sanderson MJ, Wojciechowski MF: **Improved bootstrap confidence limits in large-scale phylogenies, with an example from Neo-Astragalus (Leguminosae)**. *Systematic biology* 2000, **49**(4):671-685.

72. Soltis PS, Soltis DE: **Applying the Bootstrap in Phylogeny Reconstruction**. *Statistical Science* 2003, **18**(2):256-267.

73. Schneider TD, Stephens RM: **Sequence logos: a new way to display consensus sequences**. *Nucleic Acids Res* 1990, **18**(20):6097-6100.

74. Mason JM, Arndt KM: **Coiled coil domains: stability, specificity, and biological implications**. *Chembiochem* 2004, **5**(2):170-176.

75. Lupas A: **Coiled coils: new structures and new functions**. *Trends Biochem Sci* 1996, **21**(10):375-382.

76. Newman JR, Wolf E, Kim PS: **A computationally directed screen identifying interacting coiled coils from Saccharomyces cerevisiae**. *Proceedings of the National Academy of Sciences of the United States of America* 2000, **97**(24):13203-13208.

77. Lupas A, Van Dyke M, Stock J: **Predicting coiled coils from protein sequences**. *Science* 1991, **252**(5010):1162-1164.

78. Wolf E, Kim PS, Berger B: **MultiCoil: a program for predicting two- and three-stranded coiled coils**. *Protein Sci* 1997, **6**(6):1179-1189.

79. McIlhatton MA, Burrows JF, Donaghy PG, Chanduloy S, Johnston PG, Russell SE: **Genomic organization, complex splicing pattern and expression of a human septin gene on chromosome 17q25.3**. *Oncogene* 2001, **20**(41):5930-5939.

80. Thomarat F, Vivares CP, Gouy M: **Phylogenetic analysis of the complete genome sequence of Encephalitozoon cuniculi supports the fungal origin of microsporidia and reveals a high frequency of fast-evolving genes**. *J Mol Evol* 2004, **59**(6):780-791.

81. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res* 1997, **25**(17):3389-3402.

82. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG: **The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools**. *Nucleic Acids Res* 1997, **25**(24):4876-4882.

83. Guindon S, Gascuel O: **A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood**. *Syst Biol* 2003, **52**(5):696-704.

84. Letunic I, Copley RR, Schmidt S, Ciccarelli FD, Doerks T, Schultz J, Ponting CP, Bork P: **SMART 4.0: towards genomic data integration**. *Nucleic Acids Res* 2004, **32**(Database issue):D142-144.

85. Marchler-Bauer A, Bryant SH: **CD-Search: protein domain annotations on the fly**. *Nucleic Acids Res* 2004, **32**(Web Server issue):W327-331.

86. McGuffin LJ, Bryson K, Jones DT: **The PSIPRED protein structure prediction server**. *Bioinformatics* 2000, **16**(4):404-405.

87. Jones S, Thornton JM: **Analysis of protein-protein interaction sites using surface patches**. *J Mol Biol* 1997, **272**(1):121-132.

88. Gutell RR, Power A, Hertz GZ, Putz EJ, Stormo GD: **Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods**. *Nucleic Acids Res* 1992, **20**(21):5785-5795.

89. Tillier ER, Lui TW: **Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments**. *Bioinformatics* 2003, **19**(6):750-755.

90. Martin LC, Gloor GB, Dunn SD, Wahl LM: **Using information theory to search for co-evolving residues in proteins**. *Bioinformatics* 2005, **21**(22):4116-4124.

91. Cline MS, Karplus K, Lathrop RH, Smith TF, Rogers RG, Jr., Haussler D: **Information-theoretic dissection of pairwise contact potentials**. *Proteins* 2002, **49**(1):7-14.

92. Buck MJ, Atchley WR: **Networks of coevolving sites in structural and functional domains of serpin proteins**. *Mol Biol Evol* 2005, **22**(7):1627-1634.

93. Atchley WR, Wollenberg KR, Fitch WM, Terhalle W, Dress AW: **Correlations Among Amino Acid Sites in bHLH Protein Domains: An Information Theoretic Analysis**. In., vol. 17; 2000: 164-178.

94. Radaev S, Li S, Sun PD: **A survey of protein-protein complex crystallizations**. *Acta Crystallogr D Biol Crystallogr* 2006, **62**(Pt 6):605-612.

95. Altschul SF, Gish W, Miller W, Meyers EW, Lipman DJ: **Basic Local Alignment Search Tool**. *Journal of Molecular Biology* 1990, **215**(3):403-410.

96. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice**. *Nucleic Acids Res* 1994, **22**(22):4673-4680.

97. Pollock DD, Taylor WR, Goldman N: **Coevolving protein residues: maximum likelihood identification and relationship to structure**. *J Mol Biol* 1999, **287**(1):187-198.

98. Poon A, Chao L: **The rate of compensatory mutation in the DNA bacteriophage phiX174**. *Genetics* 2005, **170**(3):989-999.

99. Ansari S, Helms V: **Statistical analysis of predominantly transient protein-protein interfaces**. *Proteins* 2005, **61**(2):344-355.

100. Kyte J, Doolittle RF: **A simple method for displaying the hydropathic character of a protein**. *J Mol Biol* 1982, **157**(1):105-132.

101. Ostermeier C, Harrenga A, Ermler U, Michel H: **Structure at 2.7 A resolution of the Paracoccus denitrificans two-subunit cytochrome c oxidase complexed with an antibody FV fragment**. *Proc Natl Acad Sci U S A* 1997, **94**(20):10547-10553.

102. Sirajuddin M, Farkasovsky M, Hauer F, Kuhlmann D, Macara IG, Weyand M, Stark H, Wittinghofer A: **Structural insight into filament formation by mammalian septins**. *Nature* 2007, **449**(7160):311-315.

103. Hodge T, Cope MJ: **A myosin family tree**. *Journal of cell science* 2000, **113 Pt 19**:3353-3354.

104. O'Connell CB, Tyska MJ, Mooseker MS: **Myosin at work: motor adaptations for a variety of cellular functions**. *Biochimica et biophysica acta* 2007, **1773**(5):615-630.

105. McGuigan K, Phillips PC, Postlethwait JH: **Evolution of sarcomeric myosin heavy chain genes: evidence from fish**. *Molecular biology and evolution* 2004, **21**(6):1042-1056.

106. Korn ED: **Coevolution of head, neck, and tail domains of myosin heavy chains**. *Proceedings of the National Academy of Sciences of the United States of America* 2000, **97**(23):12559-12564.

107. Tolliday N, Pitcher M, Li R: **Direct evidence for a critical role of myosin II in budding yeast cytokinesis and the evolvability of new cytokinetic mechanisms in the absence of myosin II**. *Molecular biology of the cell* 2003, **14**(2):798-809.

108. Bi E, Maddox P, Lew DJ, Salmon ED, McMillan JN, Yeh E, Pringle JR: **Involvement of an actomyosin contractile ring in Saccharomyces cerevisiae cytokinesis**. *The Journal of cell biology* 1998, **142**(5):1301-1312.

109. Dobbelaere J, Barral Y: **Spatial coordination of cytokinetic events by compartmentalization of the cell cortex**. *Science (New York, NY* 2004, **305**(5682):393-396.

110. Lippincott J, Li R: **Sequential assembly of myosin II, an IQGAP-like protein, and filamentous actin to a ring structure involved in budding yeast cytokinesis**. *The Journal of cell biology* 1998, **140**(2):355-366.

111. Chang F, Peter M: **Cell biology. Formins set the record straight**. *Science (New York, NY* 2002, **297**(5581):531-532.

112. Wasserman S: **FH proteins as cytoskeletal organizers**. *Trends in Cell Biology* 1998, **8**(3):111-115.

113. Zahner JE, Harkins HA, Pringle JR: **Genetic analysis of the bipolar pattern of bud site selection in the yeast Saccharomyces cerevisiae**. *Molecular and cellular biology* 1996, **16**(4):1857-1870.

114. Kadota J, Yamamoto T, Yoshiuchi S, Bi E, Tanaka K: **Septin ring assembly requires concerted action of polarisome components, a PAK kinase Cla4p, and the actin cytoskeleton in Saccharomyces cerevisiae**. *Molecular biology of the cell* 2004, **15**(12):5329-5345.

115. Kinoshita M: **The septins**. *Genome Biol* 2003, **4**:236.

116. Leipe DD, Wolf YI, Koonin EV, Aravind L: **Classification and evolution of P-loop GTPases and related ATPases**. *J Mol Biol* 2002, **317**:41-72.