Determination of Common Motifs in Proteins Found to Interact with TriC/CCT Subunits During Large-Scale *Saccharomyces Cerevisiae* Pulldown Experiments

Christina MacDougall

Biochemistry 218: Computational Molecular Biology

Professors: Douglas Brutlag, Ph.D.
Lee Kozar, Ph.D.

Final Project, Due: 6/6/02

Outline:

Figures & Tables:

I. Background and Introduction:

In recent years, whole genome proteomic approaches have become increasingly popular tools for scientific discovery. These approaches rapidly yield vast amounts of data regarding protein levels, expression patterns, and protein-protein interactions. However, these approaches often yield such a large amount of data that analysis is often difficult and time consuming. Fortunately, the new field of bioinformatics has arisen that specifically concentrates on creating computer based approaches to analyzing large data sets, particularly those generated by proteomic and genomic approaches.

This paper focuses on the implementation of several bioinformatics techniques in an attempt to determine what molecular feature that facilitates the interaction of substrates with the Group II chaperonin TriC/CCT (TCP-1 ring complex/chaperonin containing TCP-1; TCP-1: tailless complex polypeptide). CCT is a ~900 kDa chaperonin complex made up of eight different subunits, TCP-1 and CCT2-8 that is involved in cytosolic protein folding in eukaryotes.

Up to this point no feature has become apparent which is believed to facilitate this interaction. One reason for this lack of knowledge is due to the fact that at present only the following CCT substrates have been characterized: actin and tubulin-related proteins; luciferase; G-α-transducin; the hepatitis B virus capsid protein; cyclin E; the EBNA1 viral protein; myosin; and the tumor suppressor VHL (reviewed in Dunn 2001). Due to this small number of substrates, bioinformatics approaches have yielded little information regarding common features and even when commonalties were found the significance is low due to the small data set size.

One secondary structure has putatively been implicated which may facilitate CCT-substrate interaction is the WD-repeat, a.k.a. beta-propeller. This structure has been found in variety of proteins with many varied functions (reviewed in Smith 1999). It is has so far always been found to be involved in protein complex assembly. The WD-repeat structure is made up of β-strand blades, arranged in a circular pattern to seemingly create a protein-loading platform (reviewed in Li 2001) (Figure 1). Each blade is made up of four β-strands. The number of blades, which ranges from 4 to 16, allows a somewhat artificial way to group the WD-repeat proteins into classes; five blades represents a WD5 class, etc. Based on experimental data, WD-repeat proteins have been implicated in the functioning and formation of signal transduction complexes, cell cycle regulatory complexes, apoptotic complexes, and transcriptional regulatory machinery (reviewed in Li, 2001). Based on these observations, and the knowledge that CCT interacts with β-strands in known substrates, it is highly possible that the WD-repeat may possibly be important in CCT-substrate interaction.

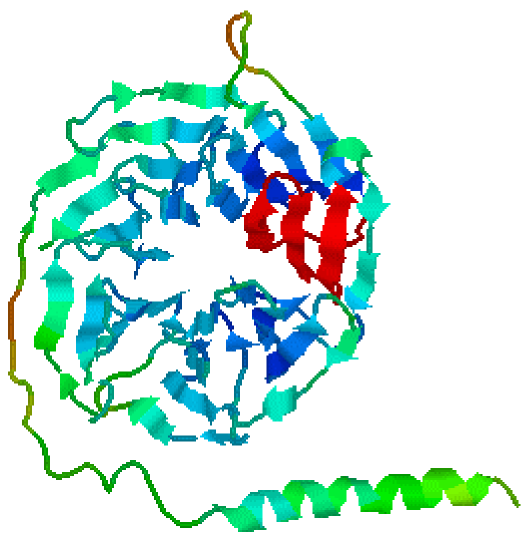

Figure 1: The beta subunit of the G protein. WD-repeat/β-propeller representative secondary structure. The red colored segment corresponds to a single WD-repeat; the above would be grouped in the WD7 Class.
(Image Obtained from http://bmerc-www.bu.edu/wdrepeat/Gb.html)

Since, as stated above, the set of known CCT binding proteins is very small, no significant bioinformatics analyses could be employed to any large extent. Fortunately, two recent papers may help to clarify the confusion regarding CCT-substrate interaction. In January 2002, two papers report data from large-scale pulldown experiments in *Saccharomyces cerevisiae*. Although each paper focuses on protein-protein interaction detection, the two methodologies employed have distinct differences. The first method employed the use of a FLAG tagged bait protein, followed by standard anti-FLAG-sepharose columns for separation of complexes out of whole yeast lysates (Ho 2002). The second method utilized the standard tandem-affinity purification (TAP) protocol (Gavin 2002). Although both methods are routinely used for pulldown experiments, they are not equal in their ability to detect interactions, especially transient ones, such as chaperonin-substrate interactions, as will be discussed later.

From the FLAG method, 80 proteins (approximately 1-2% of the yeast genome) were found to interact with CCT subunits, out of a total of 725 initial bait proteins (approximately 10% of the yeast genome) (Ho 2002). The bait proteins used were representative of multiple functional protein classes (Ho 2002). In order to facilitate the analysis, these 80 bait proteins were divided into two subgroups. The first group, Group A, are those bait proteins that had been found to interact with three or more CCT subunits, a total of 23 proteins. The second group, Group B, are those that interacted with one or two CCT subunits, a total of 57 proteins. From the TAP method, 10 proteins, Group C, were found to interact with one or two CCT subunits, out of 1739 genes that were TAP tagged (Garvin 2002). Consequently, there is only one data set from the TAP paper.

First, before beginning the bioinformatics analysis, a discussion concerning the biological relevance of the above three data sets is pertinent. The main point that needs to be addressed is the difference in the pulldown methods, FLAG vs. TAP, and why FLAG gives 80 proteins and TAP only gives 10. Most likely the difference is due to the fact that FLAG pulldowns are generally less stringent, which is excellent for detecting transient interactions, such as chaperone-substrate interactions. Consequently, FLAG pulldowns tend to contain more false positives. TAP pulldowns, on the other hand, are generally more stringent, which is excellent for eliminating background noise and false positives. While TAP pulldowns tend to have fewer false positives, transient interactions are usually lost as well. Without going into a detailed examination of the exact differences between the two protocols, which is beyond the scope of this paper, the above description fits with the data and logically explains the inconsistencies between the data sets, as we shall see below.

II. Bioinformatics Methods and Discussion:

In order to analyze sequences and structural motifs, all protein sequences were obtained. All sequences used in the following analyses were obtained from http://www.ncbi.nlm.nih.gov/ and are listed in Table 1. Although very useful, and essential to many bioinformatics approaches, the NCBI database could use some "housekeeping." Due to the fact that 90 sequences had to be obtained, the problems with NCBI were more prevalent than usual. Sequences are entered as fragments, entered under multiple accession numbers, and multiple names. For every sequence obtained, five or more sequences had to be examined in the NCBI database in order to choose the appropriate one. That wastes a great deal of time and could be easily avoided if a position were created to, at the very least, consolidate identical sequences into a single, numbered entry.

Once the sequences were obtained, general sequence alignments were performed using ClustalW (http://www.ebi.ac.uk/clustalw/). Alignments for each protein set are shown in Table 2. As one can see, there is no significant amount of alignment for the three groups of sequences. Even where ClustalW attempts to place a block, there is no real consensus sequence.

Since the sequence alignment failed, as I had believed it would, the next logical step was to look at secondary structural characteristics.

There are many programs and approaches available for examining protein secondary structure. I chose to use Prosite on the Expasy server as a first approach (http://www.expasy.ch/prosite/). After analyzing all 80 proteins using Prosite's Quickscan option, one motif appeared to a significant extent. The WD-repeat appeared in 16 out of 23 in Group A, 8 out of 57 in Group B, and none out of 10 in Group C, based on Prosite Quickscan Analysis (Table 3). No other repeat was detected to any significant extent during the analyses. Since the WD-repeat seemed to be relevant, a search was done for an analysis program specifically trained to analyze for WD-repeats. The BioMolecular Engineering Research Center (BMERC), affiliated with Boston University, Boston, MA, has a program on their server specific to WD-repeat analysis (http://bmerc-www.bu.edu/wdrepeat/). Figure 2 shows the Hidden Marko Model used as one of the initial models for developing the search algorithm used by the WD-repeat modeler.

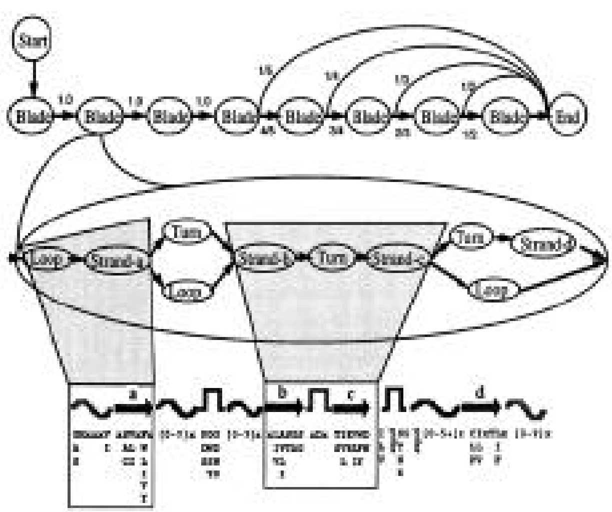

Figure 2: WD-repeat and the beta propeller HMM structure context. Displayed is a schematic graph of a beta propeller Hidden Markov model. The first line of ovals, labeled "Blade," represents a sequence of four to eight propeller blades, each composed of four beta-strands. The possible transitions, represented by the connecting arrows, are labeled with their respective probabilities. These blade states are Markov hidden states that are themselves composed of a Markov chain of hidden states displayed in second line. Here the ovals are labeled as Strand, Turn, or Loop, each of which is again a Markov chain of hidden states. The latter states are the modeled residue position states; each is assigned a set of 20 emission

Using the WD-repeat modeler, 14 out of 23 in Group A, 8 of 57 in Group B, and none out of 10 in Group C. The 14 in Group A and the 8 in Group B were the same as detected by Prosite. So, despite the fact that the WD-repeat modeler was specifically designed to detect WD-repeats, two were still not detected. However, despite the missed sequences, the modeler was still useful for the proteins in which the repeat was detected in that it was able to classify repeats into classes based on the predicted number of propeller blades. Lastly, sequences were run through the Interpro search algorithm (http://www.ebi.ac.uk/interpro/). No new WD-repeats were detected. The Interpro search detected the same set of WD-repeats as that detected by Prosite.

Next, since secondary structure appeared to be a promising avenue for analysis, whole-scale protein structure prediction was performed. Since the WD-repeats are rich in β-strands; not all WD-repeats can be detected by all algorithms, including some that may possibly have been missed by the three detection methods used. CCT is known to interact with the β-strands in some of its known substrates, I decided to employ a method to look for overall predicted β-strand percentage. The program PHD found on the PredictProtein Server (http://www.embl-heidelberg.de/predictprotein/predictprotein.html) was utilized for all whole-scale secondary structure prediction. Overall, PHD is reported to have greater than 70% accuracy for prediction of protein secondary structure (Rost 1993 and Rost 1994). However, this unfortunately still leaves a 30% error rate, almost one-third.

From the PHD outputs, a data table was assembled (Table 4). Table 4 shows all the data from all of the analyses performed for the 80 proteins. Table 4 Part A) shows the data for Group A. From the table we can see that all the proteins fall into the WD4, 5, 6, or 7 classes, based on the BMERC analysis. Since a high proportion of those proteins interacting with at least three CCT subunits contain WD-repeats (16/23 ≅ 70%), it is likely that motif plays some role in CCT substrate interaction. However, since all WD-repeats form similar structures, the specificity of the interaction is likely to lie to either the N-terminus, C-terminus, or both, of the repeat (reviewed in Li 2001). Since these flanking regions have been poorly defined and at present are a hypothetical prediction, their analysis is beyond the scope of this project. Based on the averages shown in Table 4, it appears that WD-repeat containing proteins have an average β-strand composition of approximately 12-37%; the putatively non-WD-repeat containing proteins have an average composition of 8-22%. Based on the raw data contained in the table, several of the proteins listed as not having a WD-repeat have very high β-strand compositions, indicating either a deficiency with the WD-repeat detection methods or some other β-strand based secondary structure that has yet to be characterized, and is thus undetectable via current bioinformatics methods.

Another curious pattern in the data set follows along with the hypothesis stated earlier concerning the differences between the data sets from the FLAG pulldown (Groups A and B) versus that of the TAP pulldown. One idea may be that the FLAG method did no use the same bait proteins as the TAP method; this is partly the case since the only three proteins found in the TAP pulldowns data sets were ARP2, CCR4, and SIT4. However, the more prevalent factor is that the two different protocols favor different interactions. The only two proteins found by both methods to interact with CCT subunits are ARP2 (FLAG: subunit 8; TAP: subunits 5,8) and SIT4 (FLAG: subunits 6,8; TAP: subunit 2). Due to this variety, and that fact that only one or two subunit interactions were detected with the TAP method, it appears that the more stringent TAP method favors detection of different interactions than the FLAG method. Thus, the bioinformatics results are skewed due to the skewing of interactions from the pulldown methodologies.

III. Conclusion

Using the current bioinformatics approaches, some insight has been gained as to a potential motif for interaction between CCT and potential substrates. However, when all of the ambiguities from the biological experimental differences and the bioinformatics programs, there is too much ambiguity to make any definite conclusions. Despite the fact that we have all of the pulldown data, it is difficult to determine which of these interactions are biologically relevant and which are false positives. Despite all of the various programs available for sequence analysis, no way to filter biologically relevant from background interactions, other than laboratory experiments.

Once more data concerning specific subunit interactions with specific substrates are experimentally defined, bioinformatics approaches will be able to yield more significant and specific results. For example, sequence analysis for the group of substrates that are biologically shown to interact with CCT subunit 2 can be analyzed and better similarities hopefully detected. However, despite the problems with bioinformatics approaches, more knowledge was gained concerning specifically the Group A proteins that interact with three or more CCT subunits and that the WD-repeats contained in these proteins is very likely to facilitate their interaction with TriC/CCT during protein folding reactions.

References:

Dunn AY, Melville MW, and Frydman J (2001) Review: Cellular Substrates of the Eukaryotic Chaperonin TriC/CCT. *J. of Structural Biology,* **135**:176-184.

Gavin AC, Bösche M, Krause R, Grandi P, Marzioch M, Bauer A, et al. (2002) Functional organization of the yeast proteome by systamatic analysis of protein complexes. *Nature,* **415**:141-147.

Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, et al. (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature,* **415**:180-183.

Li D and Robers R (2001) WD-repeat proteins: structure characteristics, biological function, and their involvement in human diseases. *Cell. Molec. Life Sci.*, **58**:2085-2097.

Rost B, Sander C (1993) Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.,* **232**:584-599.

Rost B and Sander C (1994) Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins*, **19**:55-72.

Smith TF, Gaitatzes C, Saxena K, and Neer EJ (1999) The WD repeat: a common architecture for diverse functions. *TIBS*, **24:**181-185.

Yu L, Gaitatzes CG, Neer EJ, and Ssmith TF (2000) Thirty-plus functional families from a single motif. *Protein Science,* **9**:2470-2476.