Using Structural Properties to Predict Behavior of Related Proteins: Does Protein Structure Influence Trypsin Miscleavage?

Micah Hamady[1], Tom Cheung[2], Henry Tufo[1], and Rob Knight[2,3]

[1]Department of Computer Science and [2]Department of Chemistry and Biochemistry, University of Colorado, Boulder, CO 80309

[3]To whom correspondence should be addressed

Running Head: Does Protein Structure Influence Trypsin Miscleavage?

Contact: rob@colorado.edu

ABSTRACT

Although resources such as the protein data bank (PDB) contain tens of thousands of protein structures, information about a protein in a particular species (e.g. human) are often unavailable. By identifying and aligning closely related sequences, we can infer structural properties of a protein in one species from the structure of its relative in another. This global homology modeling can be critical for proteomics analyses in humans, since information from mouse, fly, and even bacterial proteins can be used to make inferences about human proteins whose structure has not been directly determined.

We have constructed a database of high-quality alignments between all sequences for which proteins are deposited in PDB and the set of human proteins in IPI. The database also includes the surface area and secondary structure category of each residue in each structure, all post-translational modifications in the structures, and the mapping between coordinates in the PDB sequence and the other sequences. Crucially, the database allows a set of peptides (e.g. from shotgun proteomics) to be mapped onto the 3D structures of the closest homologous proteins in PDB.

As a case study, we report our global analysis of trypsin miscleavage in two human proteomics samples in K562 cells. We find clear differences in the propensity of trypsin to cleave in different structural contexts, although these differences may be too small to predict which additional peptides should be included for database search.

INTRODUCTION

Few of the estimated 100,000 proteins in the human proteome have been structurally characterized. This lack of information hinders proteomics analyses, since increasingly powerful techniques at the sequence level, such as shotgun proteomics [1], cannot provide access to the structures of proteins. Although resources such as Protein families database of alignments and HMMS (Pfam) [2] and SCOP (the Structural Categorization of Proteins) [3] provide useful high-level classifications of the protein family or type of fold, and therefore of the structure a protein is likely to have, these resources are difficult to use for detailed analysis at the level of individual residues and peptides.

In this paper, we develop a new database that relates structural information from proteins in PDB to closely related protein sequences in humans. Because the match criteria are extremely stringent, we can use the structure of proteins in other species to infer characteristics of the human proteins. As a demonstration of the approach, we apply this database to the problem of identifying likely trypsin miscleavage sites, which are an important problem in proteomics. However, the approach is very general, and can be used to answer many kinds of structural questions (including questions related to post-translational modifications).

Shotgun proteomics relies on the ability to identify peptide fragments from mass spectra. The fastest and most reliable method of performing peptide fragment identification is to search for matches between the actual spectra and theoretical spectra calculated from a database of sequences; two of the most popular programs for performing these types of searches are MASCOT [4] and SEQUEST [5]. The quality of

the peptide fragment search results critically depends on the size of the sequence database: larger databases have a much higher false positive rate, since the probability of seeing a chance match is greater[6]. Consequently, minimizing the size of the database is critical to ensure statistically meaningful search results.

In shotgun proteomics, proteins are digested with trypsin, a protease that cleaves protein sequences after positive residues lysine or arginine (K or R), except when they are directly followed by proline (P) [7]. Because the trypsin cleavage sites are predictable, only peptides that end with a K or R and that start immediately after the previous K or R need to be considered when searching for matches to spectra: this greatly limits the size of the peptide database, and hence the number of comparisons. It is necessary to use a protease because the ability of the mass spectrometer to generate fragment ions from a full-length protein is limited. Trypsin is used because of its high efficiency and cleavage site specificity.

Unfortunately, trypsin does not always cleave peptides consistently: trypsin can entirely miss cleavage sites, and may only affect some sites stochastically (incorrect cleavages that produce a peptide that does not end with K or R are extremely rare, although such peptides can be produced by fragmentation inside the mass spectrometer). This behavior raises a dilemma: either the database is restricted to contain only correctly cleaved peptides, which eliminates the possibility of identifying any miscleaved peptides, or the database is expanded to include all possible skipped cleavage sites, which would vastly increase the false positive rate.

Consequently, identifying factors that cause trypsin to miscleave peptides would provide a major benefit to proteomics studies, because only the peptides corresponding to

likely miscleavages would need to be added to the sequence database. In this paper, we test several properties related to the structure of proteins to test whether these structural features affect the rate of trypsin miscleavage. First, we test whether secondary structures at the cleavage site, such as alpha-helix or beta-sheet, have different abilities to resist cleavage. Second, we test whether changes in secondary structure at the cleavage site, such as from an alpha-helix to an unstructured region, affect the probability of cleavage. Third, we examine the surface area of both residues at the cleavage site, as well as the difference between them, to test whether the amount of surface area or change in surface area affects the accessibility of the trypsin molecule. We expected that cleavage sites with a smaller surface area (i.e. that are less exposed to solvent) would show higher rates of miscleavage, since these sites would be buried within the protein.

To test these hypotheses, we used a large data set of correctly cleaved peptides and peptides with skipped cleavage sites ('incorrectly cleaved peptides') from human K562 cells. K562 is a cell line that is a model for melanoma progression. The peptides were annotated with one or more International Protein Index [8] identifiers, specifying the protein sequences that contain the peptides. By mapping the correct and incorrect cleavage sites onto protein structures, we could detect any structural differences between the two types of sites. We only analyzed skipped cleavage sites, because incorrect cleavage sites (cleavage at positions other than R or K) are likely to have been generated in the mass spectrometer rather than by trypsin itself, and because the search parameters that were used to generate the data excluded non-tryptic cleavages to minimized the false positive rate.

Although over twenty-five thousand crystal structures have been solved as of June 2004, only a small fraction of these are from human proteins (and many proteins are represented by multiple structures). Fortunately, many proteins in humans form families of related sequences, and can also be clustered with related proteins in other species. Within each cluster, we aligned PDB sequences whose structures were known with IPI sequences whose structures were unknown, and thus were able to infer structural similarities. In this way, we were able to map a relatively small number of known protein structures to a large number of human proteins and thus assess the possible affects that protein structure might have on trypsin cleavage mechanism.

METHODS

Data Sources and Programming

We downloaded the 3-dimensional protein structures (that we will refer to as PDB files) from the Protein Data Bank [9] [10] and we downloaded the human reference IPI sequences (version 2.18) from [11]. We used the correctly and incorrectly cleaved peptides from human K562 cells that were characterized in a previous analysis [6]. We used PSI-BLAST [12] to locate and cluster similar proteins in a BLAST database built from the IPI and PDB sequences, and we used Clustal W [13] to align protein clusters identified by PSI-BLAST. After we identified clusters with PSI-BLAST we removed duplicate proteins from each cluster, such that every cluster contained a unique set of IPI and PDB proteins. We built the PSI-BLAST database using only a unique set of PDB sequences, such that proteins whose structure had been solved multiple times would only count once in the analysis. We used a slightly modified version of the alphasurf program

of the progeom package [14], based on the alpha shape theory [15], to compute the accessible surface area of each residue. We used the residue structure annotations from the PDB files to study the secondary structures present at each site, and we used a PostgreSQL database [16] to store and analyze much of the extracted and generated data. The majority of the source code was written using the Python programming language [17].

Trypsin Cleavage Analysis

Analysis Overview. Mapping the peptide cleavage sites onto the solved protein structures required a complex sequence of steps. Here we give a high-level overview of each step, and later we will describe each in more detail.

We began by extracting structure and sequence information for the solved protein structures from the PDB files. Then, we used this information to compute the surface area of each residue in the solved structures. Next, we built a BLAST database using human IPI sequences along with the protein sequences we extracted from the PDB files. Then we used PSI-BLAST to identify closely related clusters of proteins, and aligned these clusters using the Clustal W. We used the generated alignments to map PDB residues to corresponding IPI residues, and we used peptides from human K562 cells to find cleavage sites within the IPI sequences. Finally, we used the residue mappings in the alignments to compute statistics for the structural properties of the cleavage sites on the IPI sequences inferred from the PDB structures.

Figure 1 provides an overview of this process, broken up in four stages: protein structure analysis, cluster identification and analysis, alignment and peptide analysis, and database grid setup and search.

Analysis Detail. We now describe each step of the analysis in greater depth. We parallelized several of these steps (specifically, Steps 2, 4, 5 and 6) on the Hemisphere cluster.

1. Extract PDB protein structures. All of the PDB structure files were downloaded, decompressed, and parsed to capture the following information (for each solved protein): atom coordinates and element names, residue and strand information, post-translational residue modifications and comments, protein classification, species and tissue (if available), secondary structure information, cross-database links (if available), and the PDB identifier. In particular, this parsing provided the identity, position, and structural category of each amino acid residue. Non-protein records (i.e. synthetics, viruses, carbohydrates, and nucleic acids), incomplete, and invalid structures were discarded. Of the 25960 PDB files (as of June 15, 2004), we found 14872 valid proteins composed of 29287 individual protein strands.

2. Compute protein surface area. To compute the surface area of each residue in the PDB proteins, we used a modified version of the open-source alphasurf program from the progeom software package. The alphasurf program takes as input the type and the coordinates of each atom in a given protein strand crystal structure, and computes the total surface area of each atom accessible by a solvent with a specified radius. For this analysis, we used the radius of water (1.4 Å), since most protein molecules are found in

an aqueous environment. Several factors complicated this step, including missing and/or mis-annotated residues in the PDB records, and restrictions on acceptable input for the alphasurf program. After we computed the surface area of each atom, we then calculated both the total and the normalized total surface area of each residue, strand, and protein. Thus, from the 3D coordinates obtained in Step 1, we were able to identify the surface area of each residue within each protein.

3. Build the BLAST Database. We used formatdb from NCBI to build a BLAST database from the IPI protein sequences (downloaded as a flat file) and PDB protein strand sequences (extracted from the PDB files). This BLAST database provides the ability to search for sequences homologous to each input sequence.

4. Use PSI-BLAST to find and cluster related proteins. We ran PSI-BLAST using very stringent e and threshold values (1E-50, indicating a probability of 10^{-50} that a match at least as good would be found in the database by chance), seeded with PDB protein strands, in order to find clusters of similar IPI and PDB proteins. We parsed and post-processed the PSI-BLAST output into unique clusters to ensure that each PDB and IPI belonged to exactly one cluster. Any cluster that did not contain at least one PDB protein and one IPI protein was discarded. Further, any cluster whose IPI sequences did not contain any mapped miscleavage sites were also excluded from the analysis. (The miscleavage mapping steps are described below). We found 941 unique clusters containing 4429 unique PDB protein strands and 5745 unique IPI sequences. Thus, for each protein whose 3D structure had been experimentally determined, we were able to find all other protein sequences that closely approximated the sequence of that protein.

5. Use Clustal W to align related protein clusters. We converted the sequences found in each unique cluster into the FASTA input format used by the Clustal W multiple alignment program [13]. We then aligned each cluster with Clustal W using the BLOSUM80 scoring matrix. Thus, we were able to find the positions in each protein sequence that corresponded evolutionarily to positions in the crystal structures.

6. Create mapping between PDB and IPI residues. We processed the alignments generated by the Clustal W program and computed the mappings between each PDB residue and all corresponding (aligned) IPI residues. For this analysis, we only looked at the R or K residues and the residue immediately following an R or K residue. Additionally, each residue in the alignment had to be mapped to the correct residue number within each protein strand in the structure. The resulting mappings were stored in mapping files that we later loaded onto a distributed grid of database servers. Since there could be multiple PDB strands and multiple IPI sequences in each alignment (up to hundreds of either type), the resulting mapping files ranged from very small to very large. We used several optimizations to reduce the cubic worst case running time of our mapping algorithm. Thus, for each residue in a sequence in IPI, we could find the corresponding residue in the crystal structure of the protein it matched most closely.

7. Analyze peptides for correct and incorrect cleavage sites. We found cleavage sites using the peptides from the human K562 cells. We defined an incorrectly cleaved site as an R or K residue that was inside the peptide but not followed by a P residue (since KP and RP dipeptides are known not to cleave). Multiple R and/or K residues at the end of a peptide were not considered miscleavage sites. Correctly cleaved sites are defined as the last residue R or K residue in a peptide, and the R or K residue

immediately preceding the first residue in the peptide. The types and the indices of the cleavage sites within each peptide are stored in the database. We define the cleavage site as the peptide bond between the residues that are cleaved, typically after the C-terminus of an R or K residue and the N-terminus of the following residue. If a particular site is found to be both correctly and incorrectly cleaved, we defined this type of cleavage site to be "both" (incorrectly and correctly cleaved). The sites that were cleaved after residues other than R or K were excluded from this analysis, since few peptides containing such sites were present in our sample. Thus, we were able to identify which positions within peptides were always, sometimes, or never incorrectly cleaved.

8. Map peptides to IPI proteins. We used the mappings between each peptide and its corresponding IPI proteins that were provided byWilliam Old and Katheryn Resing (pers. comm.). We independently verified each mapping, and the stored the locations of each peptide within each IPI sequence in the database. We discarded any peptide mapping that could not be verified. Thus, we located each peptide within the protein sequence or sequences from which it came.

9. Map cleavage site residues onto cluster IPIs. We then used the mapping from the cleavage sites to peptides and the mapping from peptides to IPI sequences to compute the locations of the cleavage site residues within each IPI. We restricted the mapping to just the IPI proteins that appeared in one of the related protein clusters (identified by PSI-BLAST) since we only care about the clustered IPIs. Thus, we could locate the peptides using the coordinates of the clusters. Because the clusters contained the sequences whose 3D structures had been experimentally determined, this step allowed us to relate cleavage sites to the 3D structures.

10. Distribute data to database grid, load and index databases. The data computed in the previous steps (with the exception of the alignment mappings), that we will call reference data, was replicated across a grid of database servers using an asynchronous "one to all" replication scheme. Since the full alignment maps required a significant amount of disk space we used the database grid to load, index and analyze the cluster alignment maps in parallel. However, the amount of time and space needed for this step can be greatly reduced by excluding all residues in the alignment mapping step that are not R or K residues or that immediately follow an R or K residue (about 200x in this analysis). Alignment maps were distributed to each database node in a modified round-robin fashion to ensure equal load distribution.

11. Create denormalized tables for search. To increase search performance, we created denormalized search tables that contained only data related to the IPI and PDB sequences that were contained within the clusters.

12. Search and compute statistics for each cluster. Each database grid node performed a series of local database searches on each cluster. Each search produced information including: average surface area of each residue adjacent to a cleavage site, total counts of each type of secondary structure (helix, beta-sheet, turn, or unknown) at each residue adjacent to each cleavage site, the average surface area across a cleavage site, and the average difference in surface area between residues on either side of a cleavage site.

13. Aggregate cluster statistics and analyze results. After each database node had completed all of its local queries, we collected the statistics on the head node. We used these data to generate the graphs in Results below.

RESULTS

The trypsin miscleavage results provide a case study of the utility of our structure to homologous sequence mapping, although the technique is extremely general. The following graphs show the results of the aggregate search steps, previously described, for both average normalized surface areas as well as for secondary structure categories. Here we define the cleavage site as being between the R or K residue and the next residue.

We analyzed the relationship of miscleavage to several structural properties, including: 1) The surface area of the R or K residue before the cleavage site, 2) The surface area of the residue immediately following the cleavage site, 3) The average surface area of both residues on either side of the cleavage site, 4) The difference, or delta, between the surface area on either side of the cleavage site, 5) The types of secondary structures before and after the cleavage site, and 6) Whether the type of secondary structure changes across the cleavage site.

We used two datasets, named the Q and R sets, which were compared to a control consisting of all real and theoretical peptides that could be generated from the proteins in the IPI database. The Q and R sets were both extracts of the erytholeukemia cell line K562, grown in suspension as previously described [18]. Two samples were compared: control K562 cells and K562 cells treated with tetradecanoylphorbol acetate (TPA). Sample preparation for mass spectrometric analysis was as described [6]. Lists of peptides for control K562 cells (Q set) and K562 cells treated with TPA (R set) were generated after the data analyses which employed the IPI human protein database

(http://www.ebi.ac.uk, version 2.18, updated April 10, 2003) for identification of the accession number of the proteins for a given peptide in the set [6]. These datasets were subsequently used for this study. Although there were more peptides/proteins in the Q set than the R set, more of the proteins in the R set map to solved structures, and therefore appear in the structural analysis. Consequently, we have about twice as much structural information for the R set than we do about the Q set. This may reflect the fact that the Q set contains lower-abundance proteins, which may make identification by mass spectrometric methods less likely. The molecular functions of each set were visualized using GO-Getter [19] (Figures 7 and 8) and as expected the molecular functions are nearly identical.

Effect of residue surface area at cleavage sites on cleavage.

Figure 2 shows the effects of individual residue surface area on either side of the cleavage site on the behavior of trypsin. For the residue preceding the cleavage site, in both the Q set (Figure 2a) and the R set (Figure 2b), the distribution of surface area for the correctly cleaved sites is similar to the control. In contrast, all sites that were incorrectly cleaved at least sometimes have a greater average surface are than the correctly cleaved sites. One difference between the Q-set and the R-set is that the series with both correctly and incorrectly cleaved sites in the R-set has a much larger peak in the ~.45-.55 range. This is due to a set of proteins that are present in the R set but not in the Q set (Figure 9).

For the surface area of the residue immediately following the cleavage site, the distributions are strikingly different. In the Q set (Figure 2c), the incorrectly cleaved sites

are much more exposed than those sites that were correctly cleaved. In the R set (Figure 2d), the distributions and peaks are similar. The major difference seems to be the incorrect cleaved residues have a larger peak between ~.05 and .15 in the R set than in the Q set.

Interestingly, some individual clusters showed quite different distributions of surface areas when compared to the mean across clusters. Figure 4 shows two representative examples, in which the relative positions of the correct and incorrect peaks are reversed. These results suggest that specific rules for individual clusters of proteins may better predict trypsin miscleavage than the overall pattern.

Aggregate Average Surface Area. We measured the average surface area of the residues before and after each type of cleavage site in each of the two data sets (Figure 4). The patterns for the Q-set and the R-set were similar, although the surface areas after the cleavage sites for incorrectly cleaved peptides in the Q-set were slightly greater than those in the R-set. Sites that are incorrectly cleaved clearly have a greater average surface area both before and after the cleavage site. Interestingly, the residues before sites that are only incorrectly cleaved have a greater average surface area than those before sites that are sometimes incorrectly cleaved, but this pattern is reversed for the residues after the cleavage sites.

The following pairwise differences in means of the different cleavage types within each set and residue position were highly significant: correct vs. incorrect for the R set and correct vs. both for the Q and R sets for the residue before the cleavage site, and correct vs. incorrect for the Q and R sets and correct vs. both for the R set for the

residue after the cleavage site (unequal-variance two-sample t test, two-tailed p-values ranging from 0.0025 to 0.00002). In contrast, matched differences between the Q set and the R set were not significant (p-values ranged from 0.51 to 0.90).

Effect across cleavage sites. From the average surface area differences between the residues before and after the cleavage site shown in Figure 4, we suspected that the difference in surface area between the two residues for a given cleavage site might provide more power to discriminate correctly cleaved sites from incorrectly cleaved sites. Figure 5 shows the frequency distributions of surface area differences between the two residues flanking each of the three types of cleavage sites (data shown are for Q and R combined, but the individual results were essentially identical). The miscleaved sites tend to have a higher normalized surface area than the correct cleavage sites and the control. This effect is especially pronounced for sites which were always incorrectly cleaved. Interestingly, the incorrect cleavage sites have twin peaks outside of the main correct cleavage site peak, suggesting that the incorrect cleavage sites have a greater difference in surface area across the cleavage site than the correctly cleaved sites. In other words, when one residue is more accessible than the other, the miscleavage rate seems to increase. The general trend that the R or K residue before the cleavage site is typically more exposed than the residue after the cleavage site is also very clearly shown, since the difference between the upstream and downstream residue is nearly always positive.

Effect of secondary structure on cleavage sites. Figure 6a shows the distribution of secondary structures for each of the cleavage categories. As described

above, we extracted secondary structure information from the PDB files and counted the number of times each type of cleavage site appears within a helix, a beta-sheet, or a turn. The small number of residues that were annotated as being included in multiple secondary structure types was excluded from the analysis. Residues without secondary structure annotations were grouped under an "unstructured" category that should correspond to unstructured regions of the protein strands.

The distribution of secondary structures within each type is very similar between the Q and R sets. Correct cleavage sites tend to appear in helices slightly more frequently than incorrect cleavage sites, and that incorrect cleavage sites are slightly more frequent when they occur in unstructured regions. Beta sheets are somewhat less frequent in the R set than in the Q set. The distributions for positions following the cleavage site are the same (data not shown).

We also tested whether trypsin behaves differently at the interface between different structural types, or when a cleavage site is at the start of secondary structure. For example, a cleavage site occurring between a sheet and a helix might have a different miscleavage rate. Figure 6b shows that a cleavage site located entirely within one structural category has a slightly higher rate of miscleavage than one located at the boundary of two structural categories. This effect was consistent across both the Q and R data sets.

DISCUSSION

In our analysis of the structural properties of trypsin cleavage sites, we have shown that there are significant differences in both the surface area and the structural

categories of the correctly and incorrectly cleaved sites. However, the modes of the distributions (Figure 2) are not sufficiently well-resolved that we can accurately predict whether an individual peptide will be miscleaved based on the surface area of the positions surrounding the cleavage site. The preliminary data on individual clusters (Figure 3) suggest that a more refined model that exploits the unique properties of the peptides within each homologous cluster might provide sufficient resolution to predict miscleaved sites.

We found that the average surface area at the incorrect cleavage sites was greater than that of the correctly cleaved sites. We had expected the opposite, since we had expected the incorrectly cleaved sites to be more deeply buried inside the protein (and have a smaller normalized surface area), and that the correctly cleaved sites would be found more frequently at the surface (and have a greater normalized surface area). One possible explanation is that skipped cleavage sites often occur in 'nests' of acidic residues, presumably because the positive charge on the K or R is neutralized by the surrounding negative charges (K. Resing, pers. comm.). These charged regions would be far more likely to occur at the surface of the protein than in the hydrophobic core.

In the analysis of the secondary structure distributions, we found that cleavage sites within unstructured regions are slightly more likely to be cleaved incorrectly, whereas cleavage sites within alpha-helix structures were slightly more likely to be cleaved correctly. Further studies should reveal whether the 'acidic nests' occur predominantly in these unstructured regions.

Either or both of these analyses may have been complicated by the effects of amino acid composition, since some amino acids tend to be found more often at the

surfaces of proteins, while others tend to be buried. For example, R and K are both charged, and are expected to be found at the surface since their interaction with the dipole of water molecules is highly favorable (as opposed to uncharged, hydrophobic amino acids such as I and L). Similarly, some amino acids are more frequently found in beta sheets overall, while others are more frequently found in alpha helices. If amino acids are unevenly distributed among proteins to begin with, it may be difficult to detect structural patterns in cleavage sites that must always be formed by the same amino acids. We expect that cluster-specific analyses with higher-quality alignments (or that analyze only the best-aligned regions of the proteins rather than the whole alignments) will be able to resolve some of these issues. Additionally, improved methods for identifying peptides from spectra are likely to yield higher-quality data sets for input into this kind of analysis.

CONCLUSION

We have found that both the surface area and the secondary structure of cleavage site have a highly statistically significant affect on trypsin cleavage. The results of this analysis do not, however, suggest that surface area or secondary structure properties of particular peptides can be used to predict miscleavage sites, at least at a global level.

If miscleavage cannot be predicted even from experimentally determined 3D structures, the possibility for predicting miscleavage from computational secondary structure predictions is minimal. This leaves the original database problem unresolved: without a method for restricting the types of possible miscleavage added to the database, users are faced with the problem of either leaving all miscleaved peptides out of the database, making them impossible to identify, or including all possible miscleavages,

greatly increasing the number of false identifications. However, the discrimination provided in individual clusters may allow us to infer rules that depend on the sequence as well as the structural characteristics, which would allow prediction of miscleaved sites (and hence predictions about which miscleaved peptides to include in the database).

Our structural mapping approach may be more useful for analyzing limited proteolysis experiments. In these experiments, limited amounts of protease are added to a protein or a complex, such that only a few protease cleavages are generated during the digestion. This technique is used to identify the domain structure of large proteins, large `floppy' loops on the surface, areas of major conformational change, or protected interfaces in complexes. The ability to predict skipped cleavage sites would increase the power of these experiments, since sites that are likely to be miscleaved because of surrounding sequences would not be counted as evidence for or against particular hypotheses about the structure.

This analysis of cleavage sites demonstrates the general power of homology-based techniques, in which the characteristics of a single protein whose structure has been solved can be used to infer properties of other proteins. We expect that our database of related proteins, structures and sequences, and our ability to query experimentally determined sets of peptides against this database, will allow us to answer many other questions related to global protein expression and modification.

ACKNOWLEDGEMENTS

We thank Karen Meyer-Arendt for the list of cleaved and miscleaved peptides, and Katheryn Resing, and Natalie Ahn for discussion and advice.

Computer time on the Hemisphere Beowulf cluster (http://hemisphere.cs.colorado.edu) was provided by equipment purchased under NSF ARI Grant #CDA-9601817 and NSF sponsorship of the National Center for Atmospheric Research.

REFERENCES

[1] W. H. McDonald and J. R. Yates, 3rd, "Shotgun proteomics: integrating technologies to answer biological questions," *Curr Opin Mol Ther*, vol. 5, pp. 302-9, 2003.

[2] E. L. Sonnhammer, S. R. Eddy, E. Birney, A. Bateman, and R. Durbin, "Pfam: multiple sequence alignments and HMM-profiles of protein domains," *Nucleic Acids Res*, vol. 26, pp. 320-2, 1998.

[3] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, "SCOP: a structural classification of proteins database for the investigation of sequences and structures," *J Mol Biol*, vol. 247, pp. 536-40, 1995.

[4] M. Hirosawa, M. Hoshida, M. Ishikawa, and T. Toya, "MASCOT: multiple alignment system for protein sequences based on three-way dynamic programming," *Comput Appl Biosci*, vol. 9, pp. 161-7, 1993.

[5] Y. J. Eng J., "SEQUEST," vol. 2004. La Jolla, 1997.

[6] K. A. Resing, K. Meyer-Arendt, A. M. Mendoza, L. D. Aveline-Wolf, K. R. Jonscher, K. G. Pierce, W. M. Old, H. T. Cheung, S. Russell, J. L. Wattawa, G. R. Goehle, R. D. Knight, and N. G. Ahn, "Improving reproducibility and sensitivity in identifying human proteins by shotgun proteomics," *Anal Chem*, vol. 76, pp. 3556-68, 2004.

[7] K. A. Walsh, D. L. Kauffman, K. S. Kumar, and H. Neurath, "On the Structure and Function of Bovine Trypsinogen and Trypsin," *Proc Natl Acad Sci U S A*, vol. 51, pp. 301-8, 1964.

[8] IPI, "IPI - International Protein Index," vol. 2004, 2001.

[9] H. M. Berman, T. Battistuz, T. N. Bhat, W. F. Bluhm, P. E. Bourne, K. Burkhardt, Z. Feng, G. L. Gilliland, L. Iype, S. Jain, P. Fagan, J. Marvin, D. Padilla, V. Ravichandran, B. Schneider, N. Thanki, H. Weissig, J. D. Westbrook, and C. Zardecki, "The Protein Data Bank," *Acta Crystallogr D Biol Crystallogr*, vol. 58, pp. 899-907, 2002.

[10] ftp://ftp.rcsb.org/pub/pdb/.

[11] ftp://ftp.ebi.ac.uk/pub/databases/IPI/.

[12] I. Korf, Yandell, M., Bedell, J., *BLAST - An Essential Guide to the Basic Local Alignment Search Tool*, 1st ed. Cambridge: O'Reilly, 2003.

[13] J. D. Thompson, D. G. Higgins, and T. J. Gibson, "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice," *Nucleic Acids Res*, vol. 22, pp. 4673-80, 1994.

[14] http://csb.stanford.edu/koehl/ProShape/.

[15] J. Liang, H. Edelsbrunner, P. Fu, P. V. Sudhakar, and S. Subramaniam, "Analytical shape computation of macromolecules: I. Molecular area and volume through alpha shape," *Proteins*, vol. 33, pp. 1-17, 1998.

[16] http://www.postgresql.org.

[17] http://www.python.org.

[18] A. M. Whalen, S. C. Galasinski, P. S. Shapiro, T. S. Nahreini, and N. G. Ahn, "Megakaryocytic differentiation induced by constitutive activation of mitogen-activated protein kinase kinase," *Mol Cell Biol*, vol. 17, pp. 1947-58, 1997.
[19] M. Hamady and R. Knight, "GO-Getter: Using the Gene Ontology to Detect Differences Between Expression Data Sets," *Bioinformatics*, Submitted.

FIGURE LEGENDS

Figure 1: Overview of Trypsin Cleavage Analysis

Figure 2: Cleavage of residues surrounding the cleavage sites in the Q-set and R-set peptides. Relationship between the normalized surface area near the cleavage site (x axis) and the number of sites (y axis) for each type of site: correct cleavage (thick black line), incorrect cleavage (thick grey line), both correct and incorrect cleavage (thin black line), and control (black dashed line). The control consisted of all theoretical peptides in the sample. Data shown are for the Q-set (a and c) and the R-set (b and d), showing the residue preceding the cleavage site (a and b) and for the peptide following the cleavage site (c and d).

Figure 3: Individual Clusters. Individual cluster analysis (a and b).

Figure 4: Average surface area of residues before and after the cleavage site. Average normalized surface area of the residues before and after correctly (grey), incorrectly (black), and both correctly and incorrectly cleaved sites (white) in each of the Q and R sets.

Figure 5: Relationship between miscleavage and difference in surface area across the cleavage site. Difference in normalized surface area between the residues before and after the cleavage site (x axis) against the frequency of sites (y axis) for each of the three cleavage categories. Data shown are for the Q set and R set peptides combined.

Figure 6: Distribution of secondary structures in different cleavage site types. (a) For each of the Q set and R set, for each of the correctly cleaved, incorrectly

cleaved, and both correctly and incorrectly cleaved sites, we show the fraction of upstream residues in each of the structural categories: unstructured (light grey), alpha-helix (black), beta-sheet (white), and beta-turn (dark grey). Incorrect cleavage sites are more likely to occur in regions annotated as unstructured than in regions annotated as helices or sheets. Sheets are underrepresented in both the Q and R sets compared to the control of all possible peptides. Each group of bars sums to 1. **(b)** Effect of changes in structural category across cleavage sites. For the Q and R sets, we plot the fraction of sites that are correctly cleaved, incorrectly cleaved, separated by whether the structural categories of the two residues on each side of the site are the same or different. For example, the bars for the three cleavage types sum to 1 for the Q-set sites where the residues on both sides of the site are in the same structural category.

Figure 7: Functional comparison between Q and R sets.

Figure 1

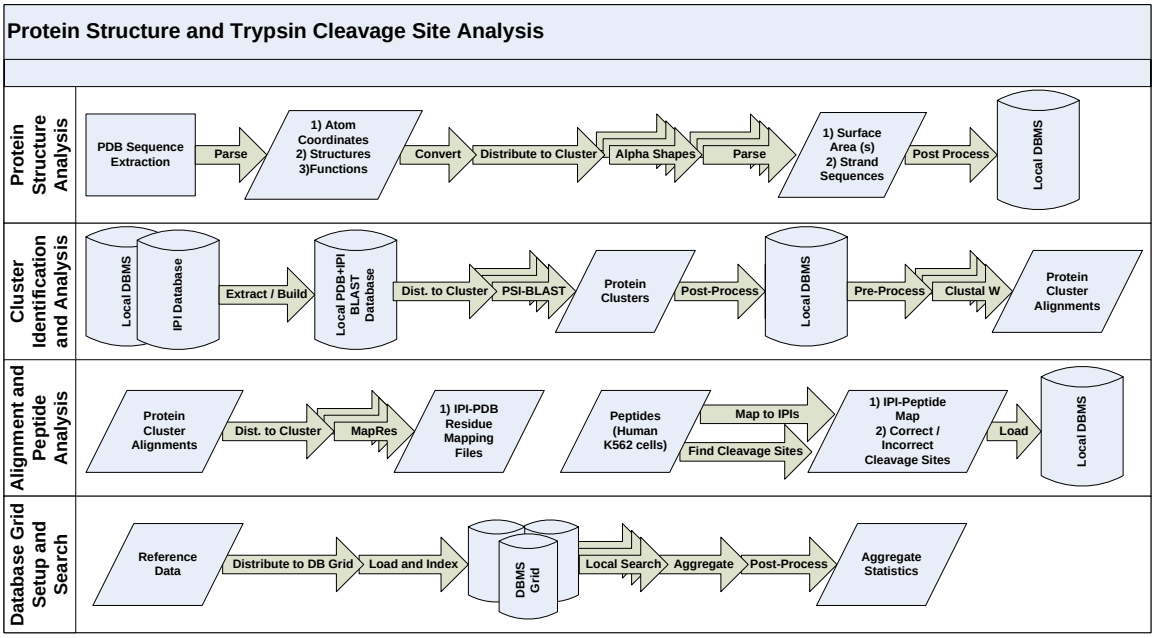
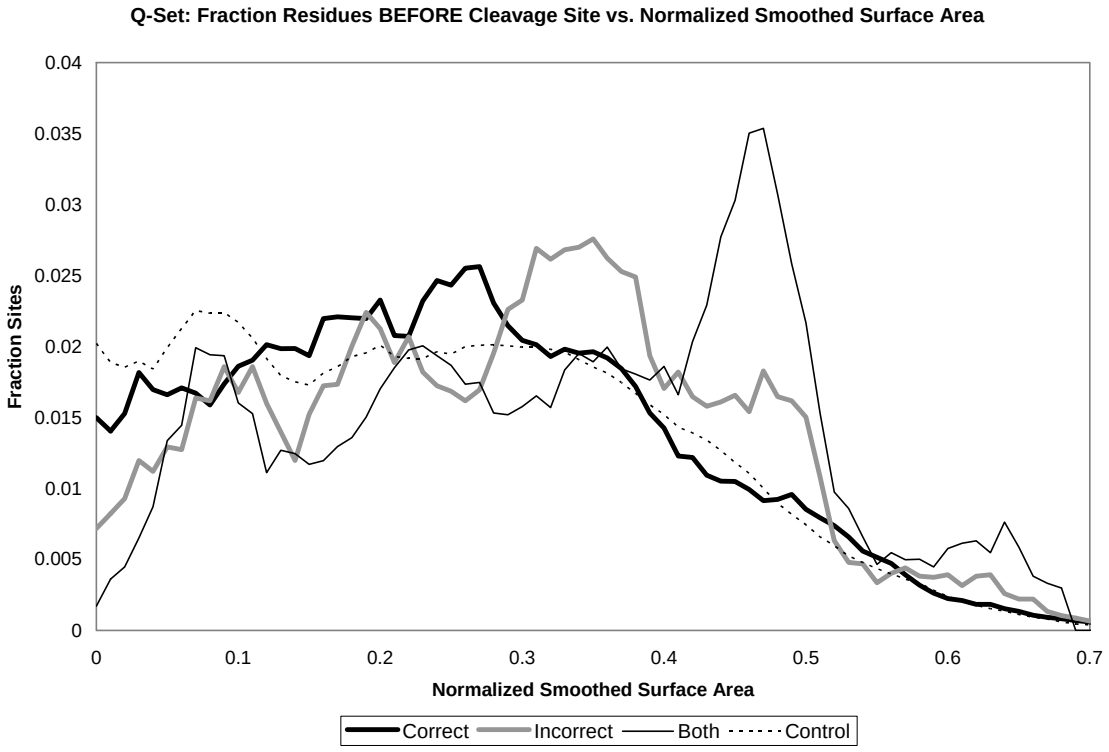
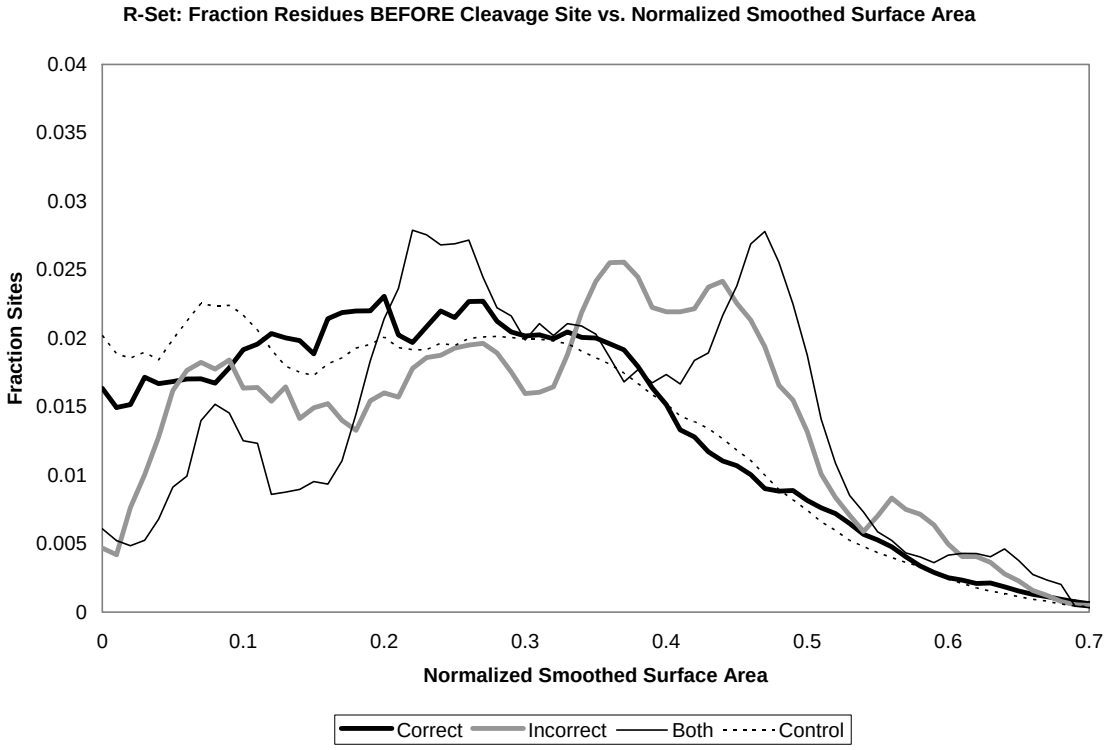

Protein Structure and Trypsin Cleavage Site Analysis
Protein Structure Analysis
PDB Sequence Extraction
Parse
1) Atom Coordinates 2) Structures 3)Functions
Convert
Distribute to Cluster
Alpha Shapes
Parse
1) Surface Area (s) 2) Strand Sequences
Post Process
Local DBMS
Cluster Identification and Analysis
Local DBMS
IPI Database
Extract / Build
Local PDB+IPI BLAST Database
Dist. to Cluster
PSI-BLAST
Protein Clusters
Post-Process
Local DBMS
Pre-Process
Clustal W
Protein Cluster Alignments
Alignment and Peptide Analysis
Protein Cluster Alignments
Dist. to Cluster
MapRes
1) IPI-PDB Residue Mapping Files
Peptides (Human K562 cells)
Map to IPIs
Find Cleavage Sites
1) IPI-Peptide Map 2) Correct / Incorrect Cleavage Sites
Load
Local DBMS
Database Grid Setup and Search
Reference Data
Distribute to DB Grid
Load and Index
DBMS Grid
Local Search
Aggregate
Post-Process
Aggregate Statistics

Figure 2

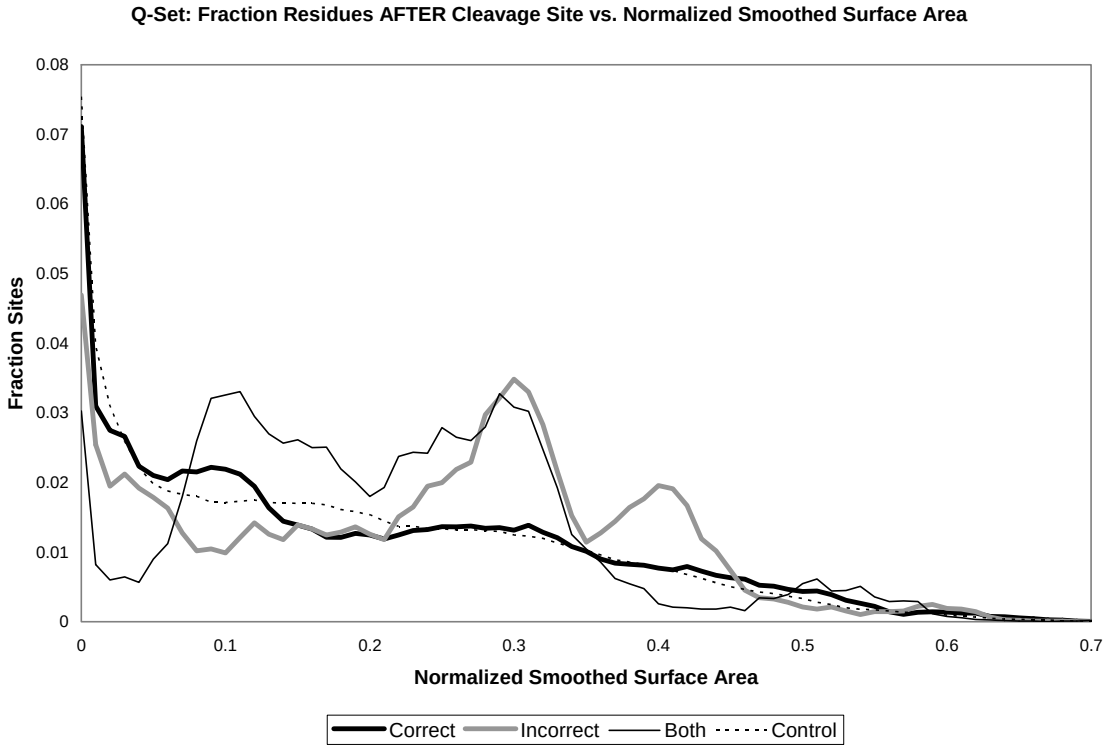
(a)


Q-Set: Fraction Residues BEFORE Cleavage Site vs. Normalized Smoothed Surface Area
Fraction Sites
0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Normalized Smoothed Surface Area
Correct
Incorrect
Both
Control

(b)


R-Set: Fraction Residues BEFORE Cleavage Site vs. Normalized Smoothed Surface Area
Fraction Sites
0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Normalized Smoothed Surface Area
Correct
Incorrect
Both
Control

(c)


Q-Set: Fraction Residues AFTER Cleavage Site vs. Normalized Smoothed Surface Area
Fraction Sites
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Normalized Smoothed Surface Area
Correct
Incorrect
Both
Control

(d)

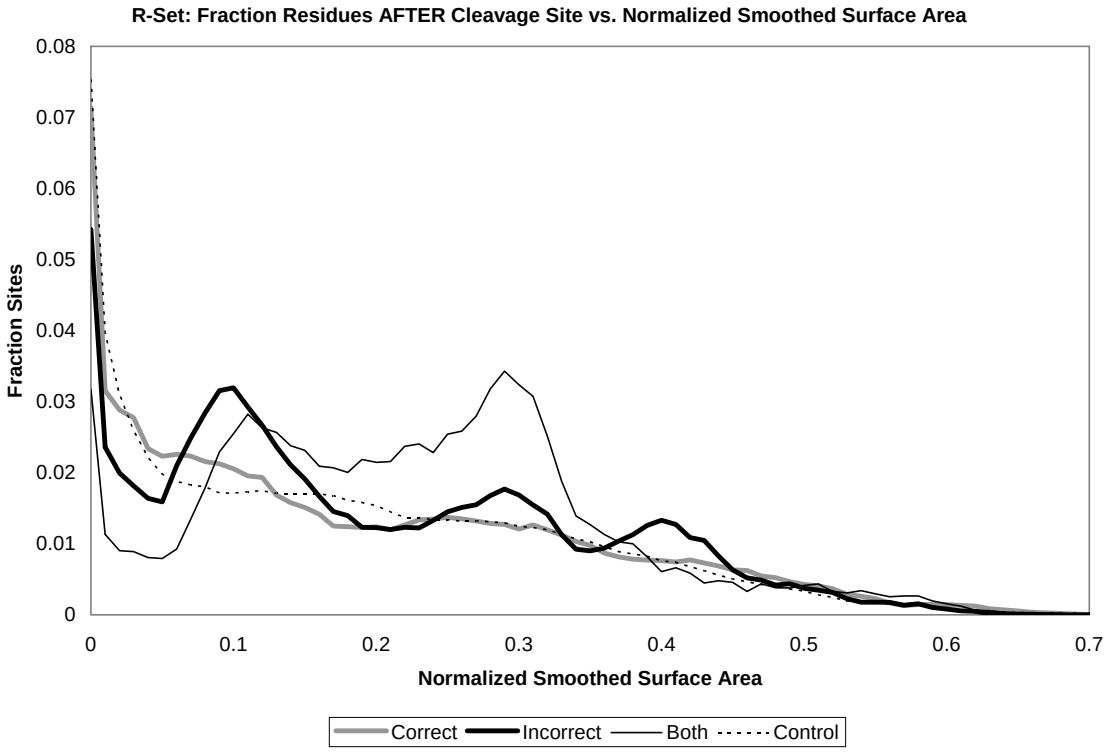
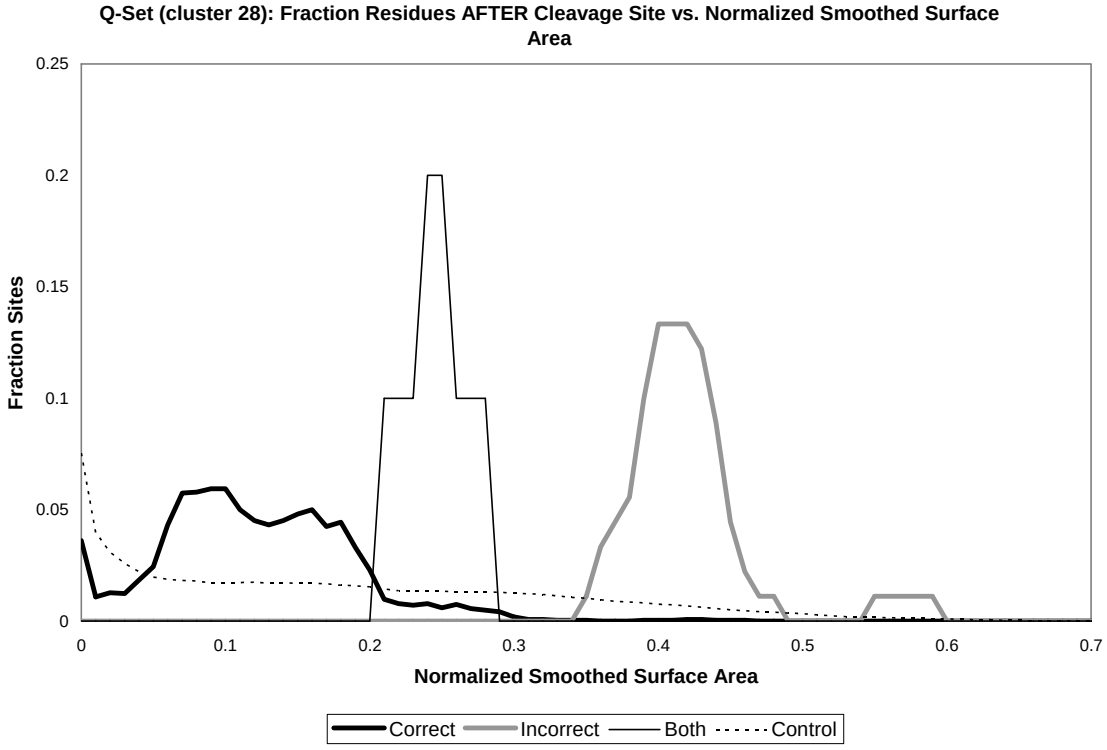

R-Set: Fraction Residues AFTER Cleavage Site vs. Normalized Smoothed Surface Area
Fraction Sites
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Normalized Smoothed Surface Area
Correct
Incorrect
Both
Control

Figure 3

(a)


Q-Set (cluster 28): Fraction Residues AFTER Cleavage Site vs. Normalized Smoothed Surface Area
Fraction Sites
0.25
0.2
0.15
0.1
0.05
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Normalized Smoothed Surface Area
Correct
Incorrect
Both
Control

(b)

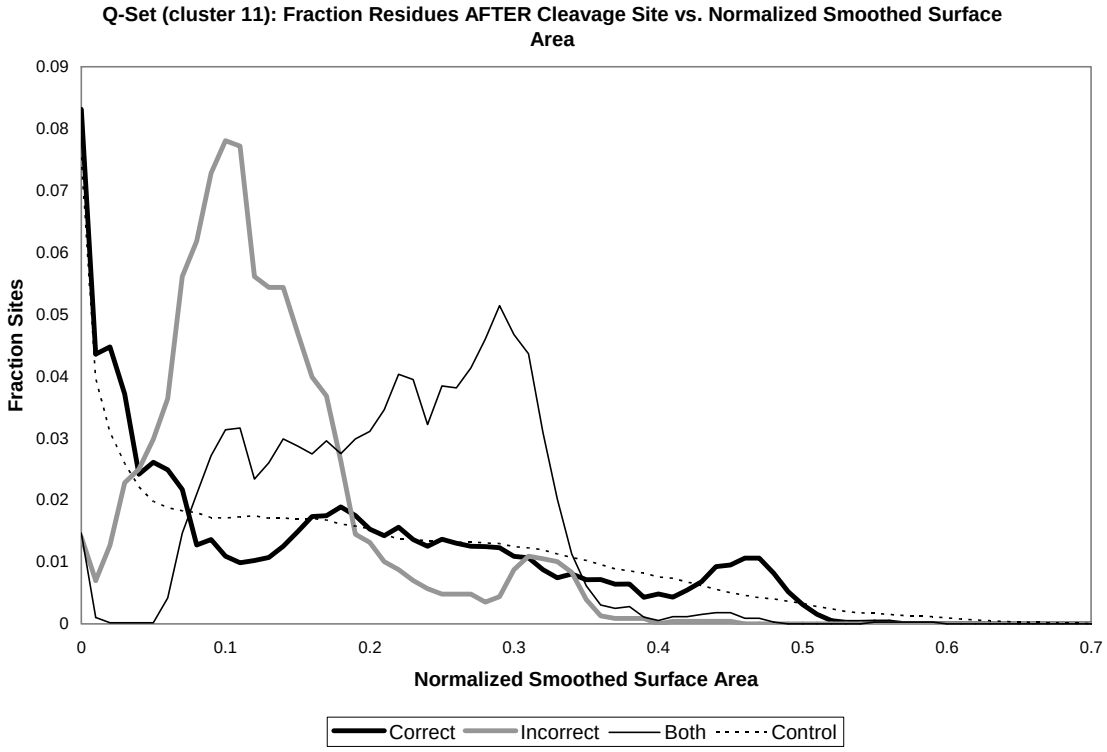

Q-Set (cluster 11): Fraction Residues AFTER Cleavage Site vs. Normalized Smoothed Surface Area
Fraction Sites
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Normalized Smoothed Surface Area
Correct
Incorrect
Both
Control

Figure 4

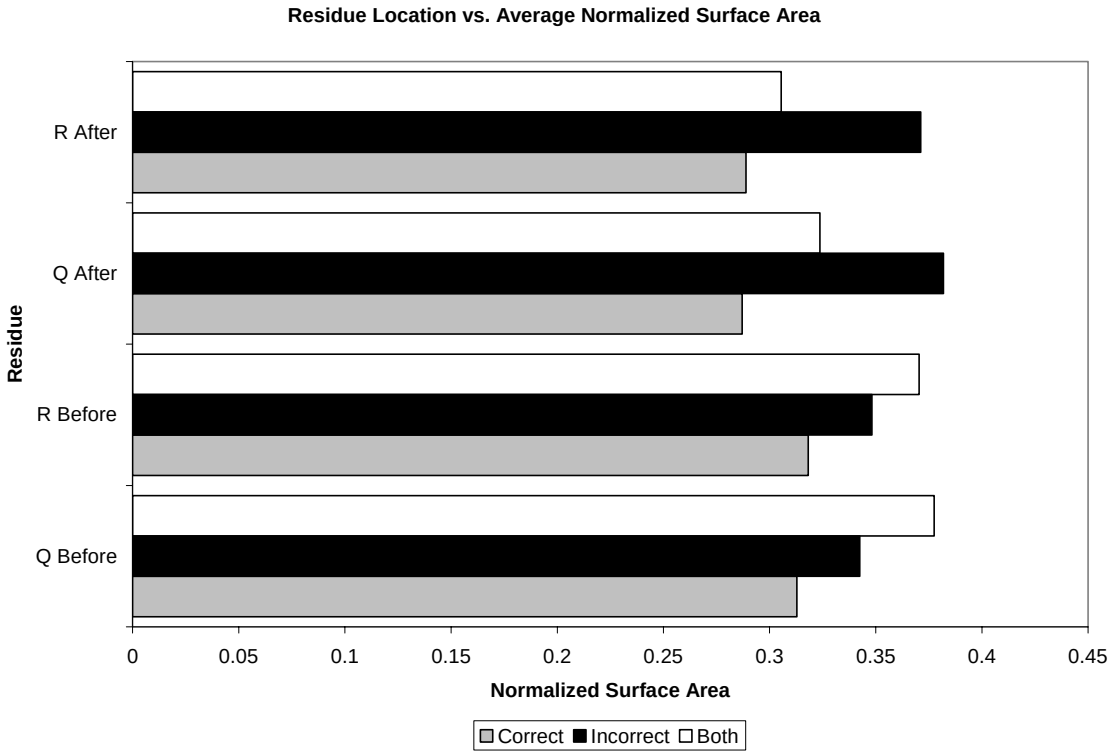

Residue Location vs. Average Normalized Surface Area
R After
Q After
R Before
Q Before
Residue
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
Normalized Surface Area
Correct
Incorrect
Both

Figure 5

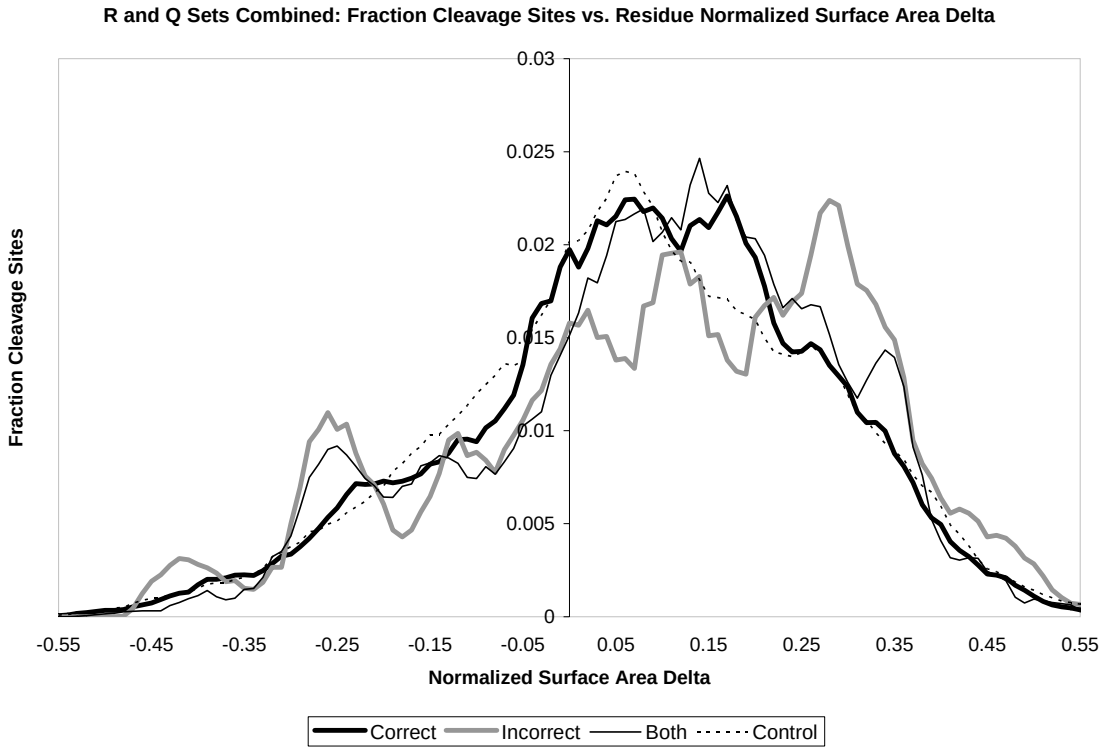
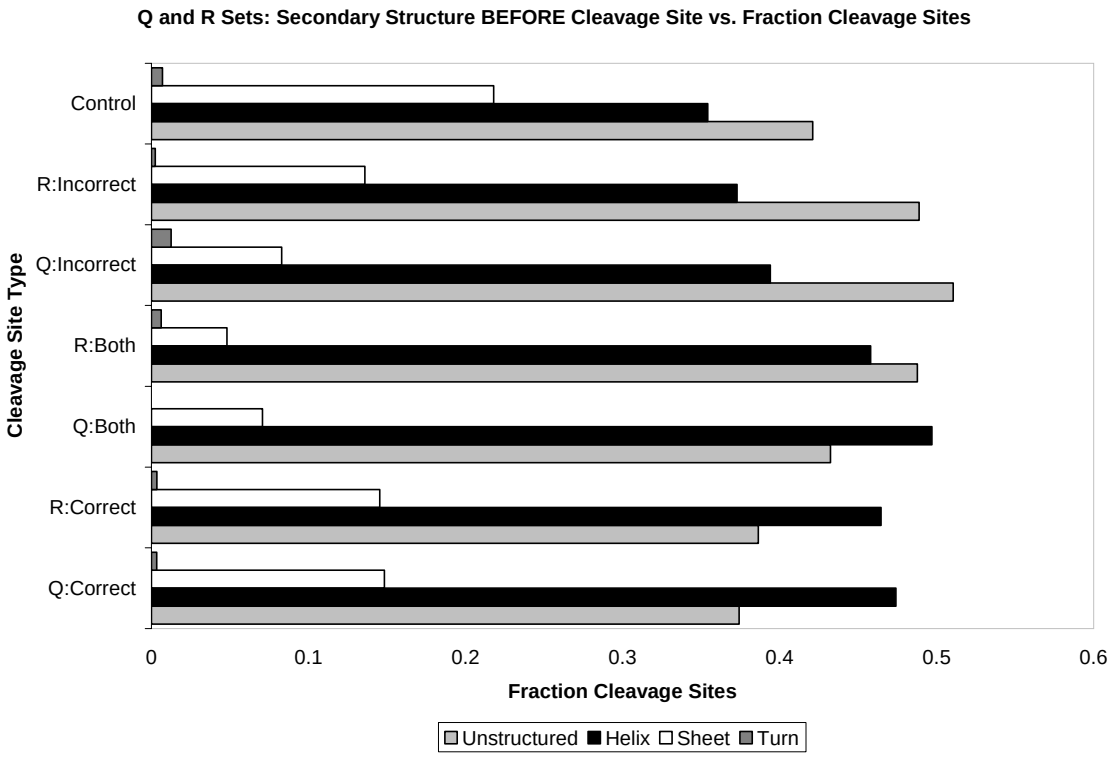

R and Q Sets Combined: Fraction Cleavage Sites vs. Residue Normalized Surface Area Delta
Fraction Cleavage Sites
0.03
0.025
0.02
0.015
0.01
0.005
0
-0.55
-0.45
-0.35
-0.25
-0.15
-0.05
0.05
0.15
0.25
0.35
0.45
0.55
Normalized Surface Area Delta
Correct
Incorrect
Both
Control

Figure 6

(a)


Q and R Sets: Secondary Structure BEFORE Cleavage Site vs. Fraction Cleavage Sites
Cleavage Site Type
Control
R:Incorrect
Q:Incorrect
R:Both
Q:Both
R:Correct
Q:Correct
0
0.1
0.2
0.3
0.4
0.5
0.6
Fraction Cleavage Sites
Unstructured
Helix
Sheet
Turn

(b)

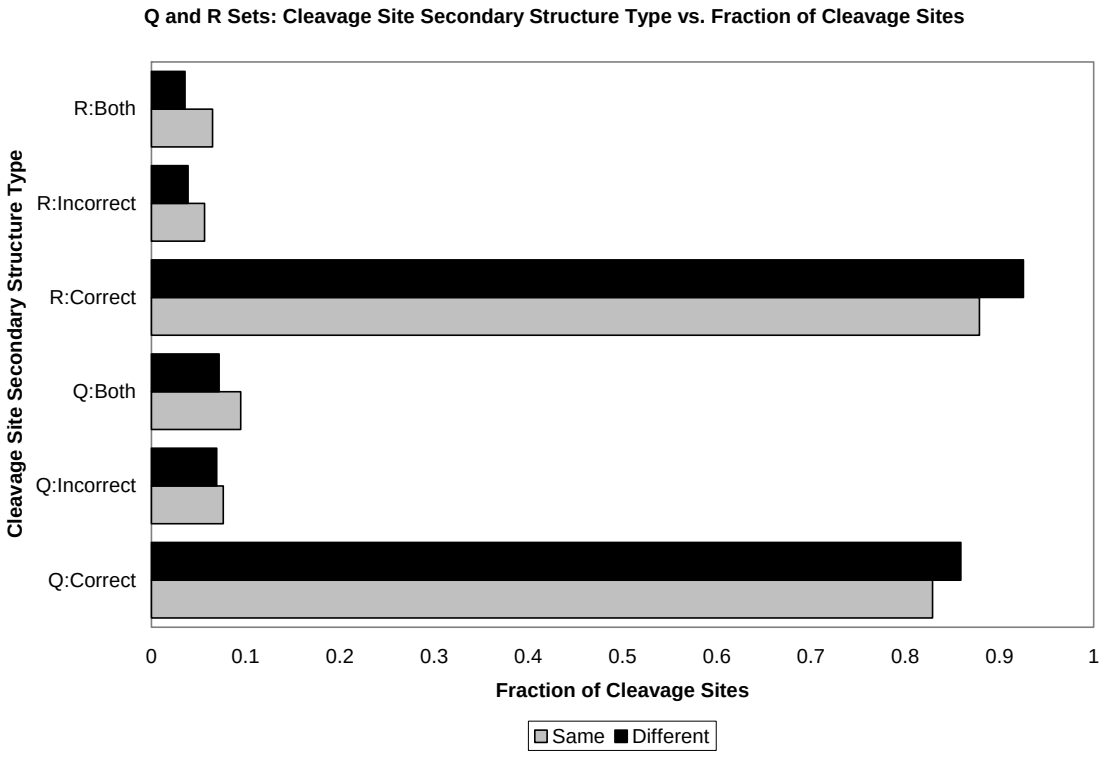
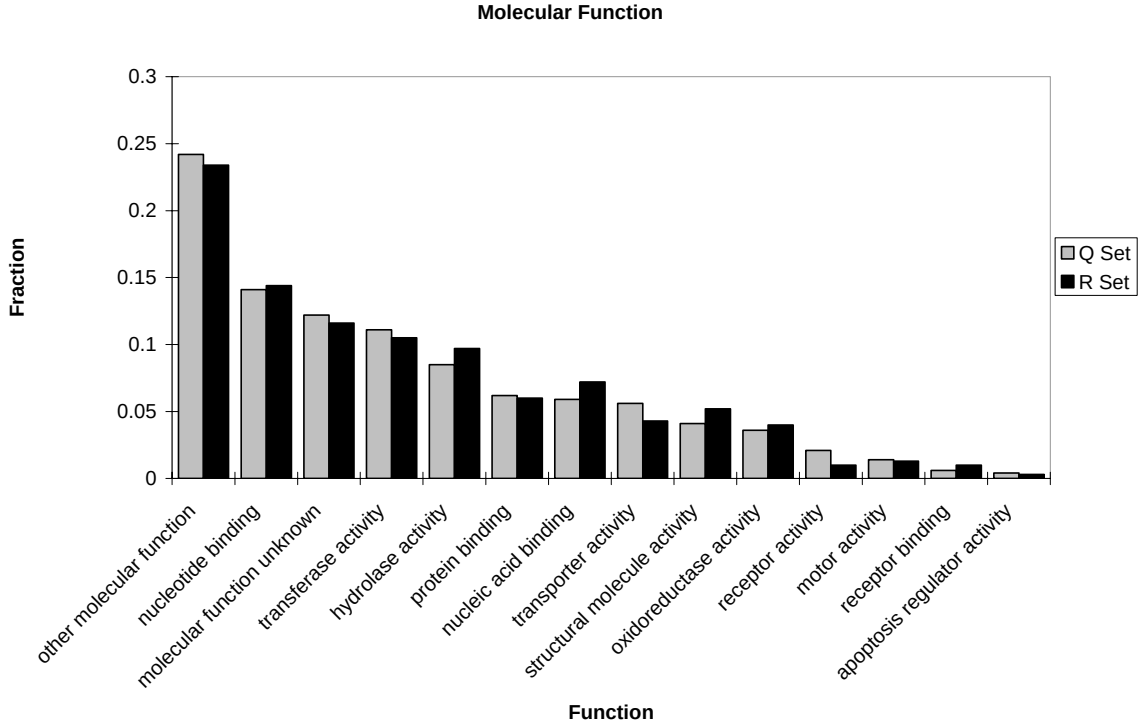

Q and R Sets: Cleavage Site Secondary Structure Type vs. Fraction of Cleavage Sites
Cleavage Site Secondary Structure Type
R:Both
R:Incorrect
R:Correct
Q:Both
Q:Incorrect
Q:Correct
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Fraction of Cleavage Sites
Same
Different

Figure 7


Molecular Function
Fraction
0.3
0.25
0.2
0.15
0.1
0.05
0
Q Set
R Set
other molecular function
nucleotide binding
molecular function unknown
transferase activity
hydrolase activity
protein binding
nucleic acid binding
transporter activity
structural molecule activity
oxidoreductase activity
receptor activity
motor activity
receptor binding
apoptosis regulator activity
Function