Sequence Motifs are Necessary but not Sufficient for Predicting Post-translational Modifications

Scott M. Carlson
Biochemistry 218
Final Paper

March 15th, 2005

Introduction

As we learn about the human genome and the protein sequences that it encodes, we are discovering that the Central Dogma of molecular biology is a useful but underpowered description of how proteins are prepared *in vivo*. Going from gene to mRNA and from mRNA to protein there are a myriad of biological interactions that complicate our understanding of the underlying systems.

A major factor complicating our understanding of biological systems is chemical modification of proteins after translation. Chemical modifications to proteins are not coded in the mRNA and they occur through protein-protein interactions. These post-translational modifications (PTM) can occur during or after a protein has folded, and they can take place in almost any subcellular region. PTMs are central in modulating almost every type of protein activity: they often control enzyme activity (Blom, 2004), change the binding affinity of protein-protein, protein-membrane, and protein-matrix interactions, bind individual peptides into larger quaternary structures, and mark proteins for destruction. Biologists studying the impact of PTMs in biological systems are challenged to catalogue the many different of PTMs, identify proteins with sites amenable to modification, and determine under what biological conditions each PTM will occur.

The first challenge in investigating PTM is the shear variety of different manners by which the amino acid sequence of a protein can be modified. The canon of molecular biology includes only twenty amino acids coded in most genomes, yet as of December 31st 2004 the RESID database of amino acid modifications contains 378 chemically distinct entries. (Garavelli, 2004) The RESID database includes only direct modifications of the amino acids, and does not include post-translational cleavage, formation of disulfide bonds, or any other PTM that modifies protein connectivity. Every one of these PTMs has a complex associated biology. PTMs demonstrate such a plethora of chemical properties that it is impossible to characterize then all using any single biochemical technique. Biological investigation of PTMs is also hindered by the fact that *in vitro* studies do not always reflect the complexity of biological systems responsible for *in vivo* regulation of PTMs.

With so many challenges facing wet-lab approaches to understanding PTMs, it has become a major area of research to understand PTMs using informatics and computational tools. Using database analysis of sequence and/or structure information allows biologists to formulate avoid wasting time and resources looking for PTMs under conditions where they are unlikely to occur. Scientists ultimately hope to have prediction tools that can scan proteome-wide databases and suggest potential PTMs. Such scans must have very high specificity to avoid having false-positive hits overwhelm the useful information.

Variability among PTMs presents a similar challenge to computational methods as it does to biological investigation. Although algorithmic approaches to PTMs need to be somewhat retooled for every situation, general methods of pattern recognition have

been applied with some success to a range of different PTMs. Although the enzymes differ among PTMs, they are all governed by the same basic physical properties: enzymes with substrate-specific binding sites interact with the target protein through their size, shape, and electrical properties, and allow some chemical reaction to occur that modifies the substrate protein. The basic problem of predicting PTMs is therefore to determine whether a protein contains sites that will be recognized by a particular enzyme. The problem of molecular recognition applies broadly to a range of biological problems (Karp, 2005) and methods developed for transcription factor binding and protein/protein interaction have been applied to this problem with variable success. The most common technique use machine-learning to recognize consensus sequences around known PTMs.

A more difficult problem, and one that has not been generally addressed, is to determine under what conditions a protein will be post-translationally modified. In addition to enzyme-substrate recognition, PTMs depend on the presence of their enzyme and often on the presence of particular chemical factors that activate that enzyme. Determining the presence of an enzyme is a problem of understanding the genetic regulatory network that governs its expression, and determining the conditions for enzyme activity requires using biochemical assays to identify the activators and repressors particular to each enzyme. Given these caveats it should be understood that this discussion will address only the issue of whether a protein *could be* a substrate for PTM, not whether it *will be* modified in any particular biological situation.

Compuational methods have several notable successes in predicting PTMs. Sequence motifs, hidden Markov models (HMM), and artificial neural networks (ANN) have all been applied with varying degrees of success. The difficulties for each of these methods are discussed later but in general they are found to predict PTMs with very high selectivity at the cost of very poor specificity (Blom, 2004). The fundamental problem is that patterns of 10-20 amino acids must be general enough to encompass the space of positive sequences with good selectivity. Such sequences occur at random in any large genome or proteome, so that ay method without very high specificity will give an unreasonable number of false-positives in any database-wide scan. Predictive annotation of a database on the scale of SwissProt requires superb specificity before predictive annotation becomes a reasonable possibility. Even specificities of > 90% on validation data sets will still produce hundreds or thousands of false positive results on huge protein databases.

The difficulties facing computational methods for predicting PTMs come out of the physical and biological mechanisms by which PTMs occur. In general, sequence based methods fail to achieve high specificity because they respond positively to matching sequences in physical contexts where a PTM cannot occur. Improvements in PTM prediction will come largely from combining a wide range of different types and of sources of information. These may be biophysical information like amino acid size and hydrophobicity, enzyme crystal structures, or evolutionary information like the degree of conservation around potential sites for PTM. For example, it has been noted that glycosylation and phosphorylation usually to occur in regions lacking secondary

structure. (Julenius, 2004, Iakoucheva, 2004) Truly robust PTM prediction algorithms will need to incorporate diverse information as well as an understanding of the genetic and other regulatory mechanisms that modulate activity biochemical pathways leading to PTMs.

PTMs are so varied and occur by so many distinct mechanisms that this review will focus on illustrative case-studies instead of providing an exhaustive catalogue. Details of all the algorithms will not be discussed, and familiarity with basic techniques of machine learning will be assumed (i.e. position-specific weight matrices, Linear Discriminant Analysis, and artificial neural networks). The first example will be N-terminal myristoylation because it is a relatively easy problem that illustrates the basic methods and challenges in PTM prediction. Phosphorylation, the second example, is a common system for enzyme regulation and it presents a set of problems that complicate the normal methods of machine learning. These examples will show how attempts to improve predictive models by incorporating physical or other information have failed or succeeded, and suggest how additional information may further improve the models.

N-terminal Myristoylation – an "easy" problem

First among the cases studies is a PTM by which an N-terminal glycine is modified by addition of myristate. Myristate is a lipid that modifies interactions between the protein and cell-membranes or hydrophobic regions of other proteins. This PTM has been shown to have involvement in virus maturation and development of cancer cells. (Bologna, 2004, Boutin, 1997) The best-studied enzyme catalyzing this type chemical reaction is Glycylpeptide N-tetradecanoyltransferase (NMT) [RESID entry AA0059].

This enzyme is an excellent example of how substrate specificity is achieved in PTMs. A cross-taxon study of NMT active sites has determined that the enzyme is selective for three regions in 17 amino acids at the N-terminal side of potential substrates. (Maurer-Stroh, JMB 2002 p.523-40) Amino acids 1-6 of the substrate must fit into an NMT binding pocket, amino acids 7-10 interact directly with the active site, and amino acids 11-17 interact with a hydrophilic region on the enzyme. Interaction between the enzyme and substrate depends very highly on the substrate's amino acid sequence only at the N-terminus.

The enzyme's strong dependence on the N-terminal sequence makes this PTM a relatively easy target for computational methods using only a short amino acid sequence of potential substrates. Furthermore, because myristoylation can only occur at terminal glycine there is no question of identifying the position of the PTM in modified proteins.

Despite the apparent simplicity of this system there are still physical limitations that can lead to false matches by any algorithm using only the primary protein structure. Myristoylation can only occur if the N-terminus of the substrate peptide is accessible to NMT. If the secondary or tertiary structure of the substrate blocks the N-terminus and/or

recognition sites then myristoylation becomes impossible regardless of how well the sequence matches the target motif.

Sequence motifs were the first approach used to identify myristoylated proteins. Based on sequences from myristoylayed proteins an early consensus sequence was found that included only the first eight amino acids, and which applied restrictions to only six of those eight locations. (Maurer-Stroh, JMB 2002 p.541-57) *In vivo* experiments and characterization of proteins in cellular extracts, along with the crystal structure of NMT, developed further restrictions on the specificity of the enzyme and led to the creation of a myristoylation pattern in the PROSITE database.[1] Bologna *et al.* applied the pattern to a set of 390 proteins with validated myristoylation and 327 proteins that are not myristoylated. (Bologna, 2004) The PROSITE pattern had sensitivity (*Sn*) of 94% (365/390) and specificity (*Sp*) of 78% (254/327).

These results indicate that proteins that are unlikely to be myristoylated if they do not match the PROSITE, but specificity of only 78% shows that the PROSITE pattern is missing much of the information relevant to myristoylation. The amino acid restrictions applied by the PROSITE pattern are almost necessary for a protein to be myristoylated but there appear to be other limiting factors that are not included in the pattern. It is unlikely that the PROSITE pattern is failing to include simple amino acid restrictions, cases where a single position is limited to a subset of amino acids, because the pattern is able to account for the first amino acid being limited to 9 out of the 20 amino acids. More likely is that the pattern misses 2nd or higher order interactions (i.e. required pairings of amino acids) or that the necessary information is simply not present in the first few amino acids. Either of these would be the case if specific secondary or tertiary structures are necessary for enzyme recognition.

Seeking to understand the mechanism by which NMT bind to substrates Maurer-Stroh *et al.* characterized the active-site of NMT by using crystal structures to compare conserved physical features among NMT in a range of species. (Maurer-Stroh, JMB 2002 p.523-40) They found that the NMT binds to 17 amino acids, as described earlier, which shows that the first six amino acids do not contain the information to predict myristoylation with high specificity. Finding a relatively unrestrictive motif this long would have been difficult without a training set much larger than was available for creation of the PROSITE pattern. Physical information about the enzyme was crucial in determining what part of the substrate amino acid sequence should be considered in predictive models of NMT. They also used this physical data to develop a classifier with a scoring function that summed a traditional probabilistic motif with score penalties for deviation from a dozen empirically determined physical rules. They did not apply to their method to a validation set, but Bologna *et al.* later showed that this is a nearly perfect classification.

Bologna *et al.* used neural networks to attempt classification based on the amino acid sequence of these 17 amino acids. (Bologna, 2004)) They used a leave-one-out procedure and achieved *Sn* 86.7% and *Sp* of 95.4% using neural networks with 320 input

[1] PROSITE pattern PDOC00008 (G-{EDRKHPFYW}-x(2)-[STAGCN]-P)

nodes (sparse coding of 16 positions with 20 inputs each), 320 2nd layer neurons, 3 3rd layer neurons, and a single output neuron. Specificity of 95.4% is very high, but it was only achieved using a somewhat over-trained classifier with 1,611 degrees of freedom.

Low selectivity and high specificity is the opposite of what is often seen when predicting PTMs and it is important to consider aspects of the algorithm that could have led to this result. It appears that this particular application of ANNs suffers from over-training even though Bologna *et al.* used a learning algorithm that restricted how precisely the network could fit itself to training data. The network fit its decision boundary too tightly to the training set, so that only sequences very similar to training data give a positive result. Specificity this high is approaching the level needed for predictive annotation of an entire database, but it still leaves much to be desired.

Both Maurer-Stroh *et al.* and Bologna *et al.* have taken classification schemes one step further by directly incorporating the physical properties of amino acids. Bologna *et al.* extended their neural network approach by creating ANNs using binary input vectors that encoded amino acid properties (large/not large, hydrophobic/not hydrophobic, etc.) for each position Their property-based ANNs used 640 input nodes and a total of 3211 independent weights. They combined the results from the original ANNs with property-based ANNs to create an aggregate classifier. Bologna *et al.* test both classifiers using a leave-one-out procedure on the same data set as the PROSITE pattern above. The algorithm by Maurer-Stroh *et al.* had *Sn* 96% and *Sp* 97% and the ANN had *Sn* 94% and *Sp* 98%. These two results are essentially identical.

Both algorithms achieve similar results by incorporating physical parameters, but they each do so in completely different manners. The rule-based approach by Maurer-Stroh takes the high selectivity of a sequence motif and augments its specificity by adding physical rules derived from outside information. It is unintuitive that adding additional degrees of freedom to the ANN improves the performance of an already over-fit classifier. It appears that physical parameters generalize more effectively than the amino acids themselves. Additional degrees of freedom allow the network to fit itself more tightly to each input vector, but because many amino acids share each physical property every training vector effectively "spans" a larger region of the input space.

That both physical methods give the same result shows that the precise method of incorporating physical data is not important. There is useful information encoded in the physical parameters that can be identified and applied by any sufficiently powerful machine-learning algorithm.

There are several issues that may be preventing even the best classifiers above from achieving better results. The first, and the most difficult to address computationally, is that the biological information may be incorrect. If a protein is myristoylated only under a limited set of conditions or only in a specific tissues then it may appear in the database as negative even when it should be classified as positive. Estimating the effect of these mistakes may be possible by examining the rate at which protein annotations change between myristoylated and not myristoylated, but a precise determination is not possible. The other factors limiting classifier performance may be a small training set,

insufficient flexibility of the models to form the decision boundary through input space, or that the first 17 amino acids simply do not contain all of the necessary information.

It is unlikely that the models are insufficiently flexibility because the ANN has a huge number of degrees of freedom. Bologna *et al.* do not report training error for their ANN, but Maurer-Stroh *et al.* report training error only slightly lower than observed on leave-one-out validation. They point out several "errors" where the annotation is in doubt so it is not clear what is causing imperfect performance on the training data.

Crystal structures are available for a limited number of the potential NMT substrates appearing the data set. Only two of the false-positives predicted by the ANN have crystal structures in the Protein Database (PDB, Berman 2000) (Aa-Conotoxin Piva and Caudina arenicola hemoglobin) so there is not enough structural information to explore in-depth what structural elements may prevent these proteins from being myristoylated even though their sequences appears to fit in the binding site. Anecdotally, the 1st and 2nd amino acids of Aa-Conotoxin Piva are both cysteine and the PDB structure shows both cysteines with disulfide bonds to other regions of the protein. These bonds lock the N-terminal region into a buried conformation, which could easily block binding to NMT (See Figure 1). The cysteines do not prevent sequence-based classifiers from calling this protein positive because the 3rd amino acid is permissive. With the hemoglobin, the N-terminus lacks 2^{o} structure and lies close to the C-terminus (~7 Å), also lacking 2^{o} structure. It may be that the flexible structure or structural interactions specific to this protein interfere with binding to NMT (rigid structures bind their targets more tightly due to entropic effects).

The structural effects that prevent NMT binding may be unique for each non-myristoylated protein that has a compatible sequence. A large set of crystal structures would be necessary to determine if they share any common structural themes. Some simple rules could be applied when a crystal structure is available. Verifying for eample that no disulfide bonds appear within some distance of the N-terminal glycine is an intuitive step that would have eliminated the false positive result with Aa-Conotoxin Piva.

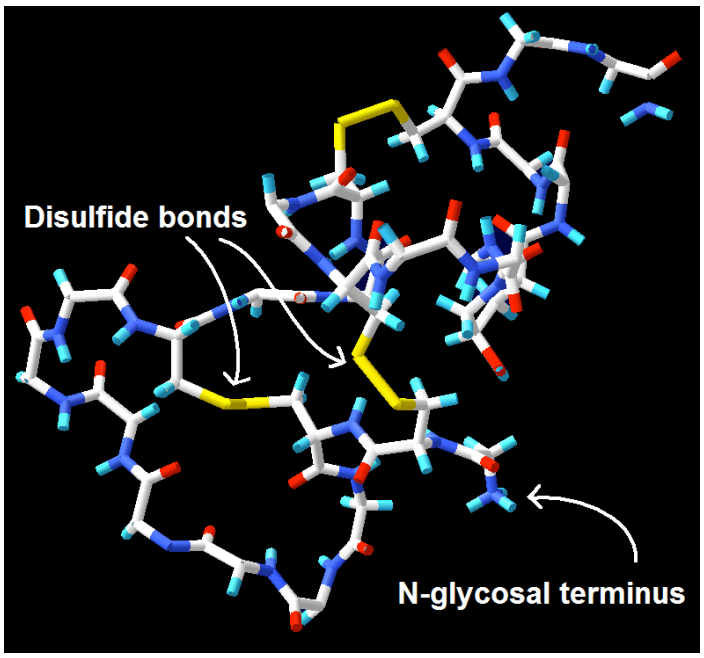

Figure 1: The PDB structure of Aa-Conotoxin Piva shows that disulfide bonds adjacent to the N-terminus are likely to interfere with sequence recognition at that position (only backbone atoms and Cys side-chains are shown)

Phosphorylation

Phosphorylation is the addition of a phosphate group to a protein, usually at the side-chain of serine, threonine, or tyrosine. This modification is a ubiquitous modification that spans every biological kingdom and which is the basic mechanism of regulation for a diverse host of enzymes. Phosphorylation reactions will often rapidly activate or deactivate an enzyme in response to regulatory signals. Phosphorylation and the reverse reaction of dephosphorylation are catalyzed by a kinase and a phosphatase respectively. There are over 500 known human kinases, each of which is active under a particular set of cellular conditions and specific to its own range of substrates. The range of different kinase specificities makes the problem of predicting sites of phosphorylation much more difficult than it was with myristoylation.

Every kinase recognizes a distinct protein substrate or set of substrates. Some are specific to a single target (the kinase for Pyruvate Dehydrogenase is associated with and targets only that protein) (Berg, 2002 Chapter 17), while others are known to phosphorylate hundreds proteins (Protein Kinase A prosphorylates over 250 substrates). (Blom, 2004) Structural studies of kinases have shown that every known kinase shares a homologous catalytic core (Berg, 2002 Chapter 10) and that they generally interact with only 12-15 amino acids around substrates' target positions, although a recent study by Iakoucheva *et al.* has found positions enriched for some amino acids with a window twice as large. (Iakoucheva, 2004) These properties make phosphorylation a good target for predictive models.

For the most broadly selective kinases there are enough positive sequences to apply machine learning techniques directly. With more specific kinases that interact with only dozens of substrates it is not feasible to develop and validate a predictive model that will span the space of $\sim 20^{12}$ possible sequences. Some methods to predict phosphorylation by more selective kinases only return hits when the same position is likely to be modified in a set of homologous proteins.

Other approaches integrate physical data and sequence information. Brinkworth *et al.* have combined kinase amino acid sequences, crystal structures, and known specificities to develop a predictive tool PREDKIN that determines the optimal substrate sequence for kinase given its amino acid sequence. (Brinkworth, 2003) Finally, Iakoucheva *et al.* have recently developed an improved predictive model that incorporates an estimate of the flexibility or disorder of possible sites of phosphorylation.

An early study in 1999 by Blom *et al.* showed that phosphorylation can predicted independent of the particular kinase involved. (Blom, 1999) They grouped the sequences of all known phosphorylation sites as serine (Ser), tyrosine (Tyr), or threonine (Thr) sites and identified the most conserved residues for each. This produces a set of empirical rules that demonstrate that there are general conserved patterns with broad applicability across the family of kinases. Methionine, for example, never occurs at the -2 position from a tyrosine phosphorylation site but occurs frequently at positions +1 and +3.

This work by Blom *et al.* led to the development of NetPhos, the oldest tool to phosphorylation prediction that is still commonly cited. NetPhos is a neural network system trained to recognize 9-11 amino acids around phosphorylation sites. The complexity of the ANN is not given in the literature but it is safe to expect that the number of weight parameters is much larger than the number of positive examples in the training set (210 for Tyr sites, 584 for Ser, 108 for Thr). The ANN implementation by Blom *et al.* achieved *Sn* 0.70 and *Sp* 0.68 on Tyr sites, *Sn* 0.89 and *Sp* 0.86 with Ser, and *Sn* 0.65 and *Sp* 0.52 on Thr on their dataset derived from the PhosphoBase database.

In the same study, Blom *et al.* attempted to use ANNs trained on three dimensional structures instead of the primary amino acid sequence. The results were unimpressive, with *Sn* 0.85 – 0.87 and *Sp* 0.37 – 0.65. These values are not directly comparable to sequence-based method above because there were very few structures available for positive proteins (12 Tyr sites for example). Structural data clearly contains useful information for predicting phosphorylation, but this result alone does not demonstrate that physical information adds anything to what can already be determined using only amino acid sequences.

All these ANNs have specificity far too poor to for predictive annotation of entire databases. This early work served to show that phosphorylation sites share common features and can be predicted in principle. More recent research has improved on this work by developing classifiers specific to individual kinases and by making novel use of other sources of information.

Both the sequence and structural methods by Blom *et al.* are limited by the fact that they are not specific to particular kinases. In addition to facing a bewildering variety of substrate specificities, this approach is confounded by the fact that both the substrate and the kinase must be present together for phosphorylation to occur. An amino acid sequence that would be phosphorylated by some kinase will appear in negative in the database if that protein/kinase pair have never been observed in a biological experiment (and in fact these sequences will be phosphorylated in *in vitro* experiments).

Yaffe *et al.* decided to avoid all the problems with creating datasets from literature by using only positive and negative results generated by testing a series of kinases on peptide library. (Yaffe et. Al, 2001) Their approach generated a large training set, but it has not been established whether peptide screens will accurately reflect phosphorylation substrates *in vivo*. (Blom, 2004) They published their results as program called ScanSite with a position specific weight-matrix for each kinase studied. Because weight matrices are well understood and computationally inexpensive, ScanSite often serves as a benchmark for other methods. It is significant that ScanSite is kinase-specific because the sequence specificity of a single kinase spans a smaller volume of the input space than the nonspecific training set from NetPhos. Having a smaller volume that gives a positive result will translate directly into fewer false-positives.

The most direct extension of NetPhos is NetPhosK, also developed by Blom *et al.* (Blom, 2004) NetPhosK makes use of the larger databases available in 2004 vs 1999 to produce kinase-specific neural networks. The ANN method is the same as in NetPhos,

but NetPhosK brings in additional information by considering the degree of conservation of a potential phosphorylated sequence across related organisms. Changes in phosphorylation will usually be associated with significant biological effects through changes in enzyme activity or other regulation. Related species are expected to have highly conserved phosphorylation sites among their homologous proteins. If only a few related species have a positively classified site then it is likely to be a false-positive. Blom *et al.* call this approach Evolutionary Stable Sites (ESS) and they expect that this approach will improve results by screening out false positives.

Blom *et al.* report data from applying NetPhosK to six different kinases with wide substrate specificity. Three of the kinases have small positive training sets (≤ 31 sequences) and because of limited training diversity the associated ANNs have *Sp* greater than 94% but with very low *Sn*. The other three kinases have reasonable positive training sets (85, 193, and 258 sequences). In all three cases *Sn* is slightly reduced from NetPhos but they have $Sp \sim 0.9$, which contrasts well to the *Sp* from NetPhos of 0.37 – 0.65. Blom *et al.* compared NetPhosK to ScanSite on a training set for the kinase PKA and found that the result from ScanSite had *Sn* 0.41 and *Sp* 0.84, while NetPhosK had *Sn* 0.84 at the same *Sp*. It was not clear from how they did not break down the results whether NetPhosK outperformed ScanSite because ANNs are more suited to discovering the decision boundary or if ESS is significantly improving specificity.

The idea of ESS is promising and the results reported by Blom *et al.* are generally positive, but it is impossible to make a strong statement on the method without more explicit results. It would be useful for Blom *et al.* to report the ROC curves with and without using ESS.

Another novel approach by Iakouchev *et al.* begins with from the observation that phosphorylation sites tend to occur in regions of disordered secondary structure. (Iakouchev, 2002) They also included a range of physical parameters in their model such as the degree of surface exposure, amino acid hydrophobicity, and side-chain bulk. Using Principle Component Analysis to select features and a variant of logistic regression as the classifier, they reported a significant improvement over NetPhos (the version that is not kinase-specific) even without incorporating any novel information.

When Iakoucheva *et al.* allowed their classifier to make sure of disordered secondary structure and other physical parameters they observed a modest improvement in the performance of their system. Defining accuracy as the average if *Sn* and *Sp* the accuracy of the classifier improved from 74.9% to 76% for Ser phosphorylation sites, 78.9% to 81.3% for Thr sites, and 81.3% to 83.3%. These improvements are all much larger than the standard error of each accuracy measurement, so there is a small but statistically significant improvement by incorporating physical parameters.

Although the classifier improvement by incorporating a disorder measure is small, it is interesting that Iakoucheva *et al.* note that over 90% of predicted phosphorylation sites show a disordered secondary structure. Disorder calculations are only minimally useful for predicting phosphorylation because they contain very little information that was not present in the sequence to begin with.

Conclusions

Algorithms to predict myristoylation have achieved nearly perfect sensitivity and specificity. In one case crystallographic studies were used to determine the binding specificity of the entire binding site of NMT and from that to augment a sequence motif with empirical rules. A completely different method based on artificial neural networks gave the same result when it was trained on a combination of sequences and physical properties of the amino acids.

Despite these impressive results, the myristoylation predictors still make mistakes that would be easy to avoid if they could incorporate a physical understanding of the enzyme/substrate interaction. The illustrative case of Aa-Conotoxin Piva shows that there is further information encoded in the full structure of the protein that is not derived from a local amino acid sequence. Since the predictors are already so powerful, it is only proteins that are somehow unusual that are going to be misclassified. As crystal and NMR structures are accumulated it will be interesting to see whether machine-learning can extract a complete description of the interactions that determine myristoylation.

Phosphorylation presents a much harder problem. Attempts to bring in additional information, both in the form of 3D structures and secondary structure prediction have met with only limited success. With such a limited set of 3D structures available it is unsurprising that machine-learning cannot extract the important information, but it is disheartening that secondary structure prediction is only slightly better than predicting phosphorylation for amino acid sequence alone. Even the best of these methods is far from perfect, so clearly there is information missing from the amino acid sequences that have been used for training.

It is not immediately clear how best to improve predictors of phosphorylation. One possible approach to the problem of small training sets for highly specific kinases would be to use the generic kinase-independent recognition pattern as a template, then modify that template for each particular kinase (for example, a Bayesian model could do a combination of the generic pattern with observed sequences from some particular kinase weighted by the number of available observations). Another approach could be to create boundaries in the decision space that are constrained by simulated docking of sequence on the boundary to crystal structures for well-characterized kinases.

Every type of post-translational modification is essentially the same problem of determining protein/protein binding interactions, so the same methods used for myristroylation and phosphorylation can be usefully applied to type of PTM. In every case however the same problem of specificity will be encountered. Normal amino acid motifs only simply do not have enough information content to specify binding sites with high enough specificity for database-wide scans without a high rate of false positives. The information encoded sequence far from the immediate binding region of each substrate is very diffuse, and cannot be extracted by machine learning without a huge number of examples. A better approach will be to understand the physical basis of each substrate/enzyme interaction and use either predicted or experimental substrate structures

as input to classifiers. As protein structure prediction improves it will become more feasible to do structure prediction on entire databases. Because it is likely that some proteins structures will be predicted incorrectly this approach may lead to a problem of high specificity and low sensitivity. It is more likely that a protein would be incorrectly folded into structure that does not bind to a PTM enzyme then that it would accidently be predicted to fold into a binding conformation (the space of binding conformations is much smaller than the space of nonbinding conformations). A really complete method for predicting post-translational modification will always be restricted to those enzymes and substrates for which complete structures have been determined experimentally.

References

1. Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The Protein Data Bank. *Nuc. Acid. Res.* **2000**, *28*, 235-242.

2. Berg, J.M.; Tymoczko, J.L.; Stryer, L. Biochemistry 5th ed. W.H. Freeman and Company. New York, **2002**.

3. Blom, N.; Gammeltoft, S.; Brunak, S. "Sequence and Structure-based Prediction of Eukaryotic Protein Phosphorylation Sites." *J. Mol. Biol.* **1999**, *294*, 1351-1362.

4. Blom, N.; Sicheritz-Pontén, T.; Gupta, R.; Gammeltoft, S.; Brunak, S. "Predicion of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence." *Proteomics*. **2004**, *4*, 1633-1649.

5. Bologna, G; Yvon, C.; Duvaud, S.; Veuthey, A. "N-terminal myristoylation predictions by ensemble of neural networks." *Proteomics*. **2004**, *4*, 1626-1632.

6. Boutin, J.A. "Mysritoylation." *Cell. Signal*. **1997**, *9*, 15-35.

7. Brinkworth, R.I.; Breinl, R.A.; Kobe, B. "Structural basis and prediction of substrate specificity in protein serine/threonine kinases." *PNAS*. **2003**, *100*, 74-79.

8. Eisenhaber, B; Eisenhaber F.; Maurer-Stroh S.; Neuberger G. "Prediction of sequence signals for lipid post-translational modifications: Insights from case ." *Proteomics*. **2004**, *4*, 1614-1625.

9. Garavelli, J.S. "The RESID Database of Protein Modifications as a resource and annotation tool." *Proteomics*. **2004**, *4*, 1527-1533.

10. Iakoucheva, L.M.; Radivojac, P.; Brown C.J.; O'Connor T.R.; Sikes, J.G.; Obradovic, Z.; Dunker, A.K. "The importance of intrinsic disorder for protein phosphorylation." *Nuc. Acids. Res.* **2004**, *32*, 1037-1049.

11. Julenius, K; Mølgaard, A; Gupta, R; Brunak, S. "Prediction, conservation analysis, and structural characterization of mammalian mucin-type O-glycosylation sites." *Glycobio*. **2004**, *15*, 153-164.

12. Karp, Richard. University of California Berkeley. Personal communication, **3/15/2005**.

13. Mann, M.; Ong, S.; Grømborg, M.; Steen, H.; Jensen, O.N.; Pandey, A. "Analysis of protein phosphorylayion using mass spectrometry: deciphering the phosphoproteome." *TRENDS Biotech*. **2002**, *20*, 261-268.

14. Maurer-Stroh, S.; Eisenhaber, B.; Eisenhaber, F. "N-terminal N-Myristoylatino of Proteins: Refinement of the Sequence Motif and its Taxon-specific Differences." *J. Mol. Biol.* **2002**, *317*, 523-540.

15. Maurer-Stroh, S.; Eisenhaber, B.; Eisenhaber, F. "N-terminal N-Myristoylatino of Proteins: Predicion of Substrate Proteins from Amino Acid Sequence." *J. Mol. Biol.* **2002**, *317*, 541-557.

16. Monigatti, F.; Gasteiger, E.; Bairoch, A.; Jung, E. "The Sulfinator: predicting tyrosine sulfonation sites in protein sequences." *Bioinformatics*. **2002**, *15*, 769-770.
Yaffe, M.B.; Leparc, G.C.; Lai, J.; Obata, T.; Volinia, S.; Cantley, L.C. "A motif-based profile scanning approach for genome-wide prediction of signaling pathways." *Nat. Biotechnol*. **2001**, *19*, 348-353.