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