

Intrinsic disorder and protein modifications: building an SVM predictor for methylation

Kenneth M. Daily,¹ Predrag Radivojac,^{1*} A. Keith Dunker^{2*}

1) School of Informatics, Indiana University, Bloomington, IN 47408

2) Center for Computational Biology and Bioinformatics, School of Medicine, Indiana University, Indianapolis, IN 46202

*Corresponding authors: predrag@indiana.edu; kedunker@iupui.edu

Abstract-Post-translational protein modifications play an important role in many protein pathways and interactions. It has been hypothesized that modifications to proteins occur in regions that are easily accessible, and many have been determined to be located within intrinsically disordered regions. However, identifying precise locations of protein modifications involves expensive and time consuming laboratory work. Thus, automated identification of these sites is helpful. This paper studies methylated proteins and describes methods of building a predictor for arginine and lysine methylation sites using support vector machines. Our results indicate that, based on current data, both arginine and lysine methylation sites are likely to be intrinsically disordered and that the accuracies of methylation site predictions are high enough to be useful for protein screening and for testing biological hypotheses.

Availability: www.informatics.indiana.edu/MethylationPredictor

I. INTRODUCTION

A. Protein Methylation

Protein methylation was discovered more than 35 years ago [1] and, together with other post-translational modifications, it offers great functional diversity to the primary sequence of a protein. However, not nearly as much is known about the processes or implications of protein methylation as is about other post-translational modifications, such as phosphorylation. Methylation can occur at many residues, including arginine, lysine, histidine, alanine, proline, aspartic and glutamic acid, and glutamine. Nitrogen atoms, either backbone or sidechain, are methylated (N-methylation) in arginine, lysine, histidine, alanine, proline, and glutamine, and oxygen atoms are methylated in glutamic and aspartic acid residues [3]. These additions are carried out by a protein family called methyltransferases, which use S-adenosylmethionine as a substrate to transfer a methyl group [4]. Originally, it was thought that methylation is not a reversible event [5]. However, recent research indicates reversibility, as supported by the enzyme LSD1 that possesses lysine demethylase activity and can remove methyl groups from lysine 4 of histone protein H3 [6]. In this study, we are focusing on arginine and lysine residues only, for which mechanisms of methylation are best understood and for which the data are publicly available.

Arginine methylation involves the addition of a methyl group to nitrogens within an arginine in a polypeptide. Three main forms of arginine methylation have been identified: monomethylarginine (N^ε-monomethylarginine), symmetric dimethylarginine (N^ε, N^{ε'}-dimethylarginine), and asymmetric dimethylarginine (N^ε, N^ε-dimethylarginine). Eight mammalian protein arginine methyltransferases (PRMTs) have been iden-

tified, six of which have been shown to transfer a methyl group to the guanidino nitrogen of arginine from S-adenosylmethionine [7]. There are two types of PRMTs, both of which form N^ε-monomethylarginine. The type I PRMTs (PRMT1, PRMT3, PRMT4, and PRMT6) produce asymmetric dimethylarginine, while the type II PRMTs (PRMT5 and PRMT7) produce symmetric dimethylarginine. There has been no documented activity for PRMT2 or PRMT8 [7].

PRMTs are found in many different tissues and generate specificity by alternative splicing [8]. Proteins with glycine and arginine-rich regions are often targets for PRMTs [9]. Methylation of arginines has been identified in roles in transcriptional regulation, RNA processing, signal transduction, DNA repair, cell-type differentiation, genome stability and cancer [7, 10]. Although it is generally thought that PRMTs are very specific, evidence exists showing the same arginine residues in a substrate protein being methylated by both type I and type II PRMTs. In addition, SmB and SmB', which are core small nuclear ribonucleoproteins, have been shown to be symmetrically dimethylated by PRMT5 [11] as well as asymmetrically methylated by CARM1 [12]. Some proteins found in SWISS-PROT [13] that are known to contain methylated arginines are PABP2, which is involved in the addition of a poly(A) tail to mRNA precursors, SFPQ, a DNA- and RNA-binding protein, and the oncogene FUS, a DNA-binding protein.

Lysine residues are methylated via a similar mechanism. A cofactor called S-adenosylmethionine is used by a class of proteins called histone methyltransferases (HMTs) to transfer a methyl group to the lysine residue. Three types of lysine methylation have been discovered: monomethyllysine, N-dimethyllysine, and N-trimethyllysine [3]. The first discovery of histone methylation occurred in the 1960s [14] and since then many studies have focused on the methylation of specific lysines in the tails of histones. These methylation sites are well known [15] and specific, indicating that methylation is not a random event [16]. Some non-histone proteins, such as p53 - a tumor suppressor protein, have also been shown to contain methylated lysines [17]. Other proteins found in SWISS-PROT with known methylated lysine residues include cytochrome C, calmodulin, and the large subunit of rubisco.

B. Intrinsically Disordered Proteins

The functional annotation of proteins known to be methylated indicates that many of these proteins are intrinsically disordered. This class of proteins is characterized by unstable tertiary structure under (putatively) physiological conditions [18-20]. These proteins sample their structures from ensembles of conformations, either in part or along their entire

lengths, and typically explore wide and irregular motions. Intrinsically disordered proteins exist both *in vitro* and *in vivo* [21, 22] and are shown to carry out important biological functions [23]. Dunker et al. [24] classified the functions of disordered proteins into four broad functional sets: (i) molecular recognition; (ii) molecular assembly/disassembly; (iii) protein modification; and (iv) entropic chain activities. Tompa [25] used slightly different notation by splitting molecular recognition into scavengers and effectors and adding chaperones [26]. Thus, disordered protein regions are typically involved in regulatory activities, signaling or control, while ordered regions, those with stable 3-D conformation, are typically involved in enzymatic activities. It has been suggested that there are two parallel structure-function paradigms in the protein world: sequence \rightarrow 3-D structure \rightarrow function for enzymes, and sequence \rightarrow disordered ensemble \rightarrow function for signaling and regulatory proteins and regions [27].

C. Outline of the Study

To learn about the methylation of lysines and arginines in proteins, we investigated 107 proteins with experimentally determined post-translational lysine methylation sites and 41 proteins with post-translational arginine methylation sites. In order to statistically associate experimentally verified methylation sites with intrinsically disordered regions, we first examined the differences in the local amino acid compositions and various physicochemical properties around the methylated sites. Then, based on distinctive features between methylated and non-methylated residues, we developed a new methylation site predictor using support vector machines (SVMs). Our model reached a balanced-sample accuracy of 77.9% and 63.1% for arginine and lysine methylation, respectively. Similarly, the area under the receiver operating characteristic (ROC) curve, as explained below, was estimated to be 85.0% for arginine and 66.4% for lysine.

To our knowledge the first published predictor of methylation sites was constructed by Plewczynski et al. within their AutoMotif Server [28]. Their method involves constructing regular expressions from experimentally verified functional sites in proteins from SWISS-PROT and creating fragments of 9 residues. The fragments are then projected to a multidimensional space of features, including orthogonal vectors, position specific features using the BLOSUM62 substitution matrix, a normalized sequence preference for an amino acid found at a certain position, as well as a real-valued ratio of preferences for a certain amino acid in each position of the positives versus the negatives. Their model was constructed using short sequence fragments only and the SWISS-PROT database, while the negative examples were selected randomly from the remaining set of proteins.

Our predictor of methylation sites was developed using a similar discriminative approach with a significantly expanded set of features and different assumptions on the negative examples. We try to learn rules for methylation in an automated fashion, without relying on high sequence similarity present in the datasets available. We utilized our biological hypothesis that methylation sites are preferentially located within intrinsically disordered regions and used already developed predictors of disorder and flexibility in order to improve classification accuracy. The results of our study

strongly indicate that many methylated sites occur within flexible protein regions and that they can be predicted with satisfactory and useful accuracy.

II. MATERIALS AND METHODS

A. Datasets

Candidate proteins were collected from SWISS-PROT, version 45, for both lysine and arginine methylation sites. From the SWISS-PROT data, all records containing the keyword methylation were first separated. A Perl script utilizing BioPerl then searched for specific terms related to these methylated proteins. Positive (methylated) sites were found by examining the MOD_RES field in each SWISS-PROT record. All records containing methyllysine or methylarginine in a MOD_RES field were collected, excluding those marked by “probable,” “possible,” “potential,” and “by similarity,” which are non-experimentally determined sites. However, the “probable” sites are partially experimentally determined, and may be included at a later date. Differences in the types of methylation (mono-, di- (symmetric) and di- (asymmetric) for arginine, and mono-, di-, and tri- for lysine) were not taken into account in data collection. The control dataset of non-methylated sites was extracted from the same proteins and includes all arginine and lysine residues not marked as methylated. The datasets were composed of 25-residue fragments, 12 residues before and after the arginine residue. However, some shorter fragments are included in the datasets that were close to the ends of the protein. The paper by Ong et al. [29] was also used as a source for arginine methylation sites.

For the lysine dataset, we collected a total of 107 proteins with 213 positive sites and 1943 negative sites. For the arginine data we had 41 proteins from SWISS-PROT and Ong et al. and a positive dataset with 116 sites and the negative set with 1315 sites. We assumed that the experimentally determined modification sites were correct and removed negative fragments with a similarity to any positive fragment of greater than 30%. In total, the arginine dataset yielded 883 fragments and the lysine dataset 1703 for negative non-redundant fragments. After removing redundant fragments within each positive and negative dataset, defined as those with a similarity of greater than or equal to 40%, the datasets contained 84 positives and 757 negatives for arginine, and 64 positives and 833 negatives for lysine.

B. Statistical Tests

Standard t-test of statistical significance was performed to determine whether the positive and negative sequence fragments around all available arginines and lysines were actually different, not just by chance. At each position around the arginine or lysine residue, we compared means between the positive and negative datasets for each amino acid and each position. Our output of this is an estimate of the p-value, or significance level after which we would reject the hypothesis that the two datasets are coming from the same distribution. In all experiments we used a p-value cutoff of 0.05 to indicate statistical significance.

We determined which amino acids were enriched and depleted at each position by using the difference in frequency

between the positive and negative datasets. Enriched residues are those that occur more frequently in the positive set than the negative set, and the depleted ones occur more in the negative than the positive set.

C. Data Representation and Predictor Construction

A dataset of features for our proteins was constructed using Matlab. For constructing features, fragments shorter than length 25 were included. We used amino acid frequencies over multiple windows (sizes 3, 7, 13, 19, and 25) around the arginine/lysine. Real-valued features were used for aromatic content (W, F, and Y), a flexibility scale by Vihinen et al. [30], net charge ($n_K + n_R - n_D - n_E$, where n_X is the number of residue X in a window), hydrophobic moment [31], sequence complexity [32] and beta entropy [33]. These were averaged within the same window sizes as used for the amino acid frequencies. We also used several predictors of structural disorder including VL2 [34], VL3 [35], and also a B-factor predictor [36]. Note that the VL2 model contains four different predictors: three specialized (VL2-V, VL2-C, and VL2-S) and a general one, denoted as VL2 [34]. Windows of sizes 1, 7, and 11 were used for the predictor features, and the mean and maximum of the features were taken for each window. Position specific scoring matrices generated by PSI-BLAST [37] were used to model evolutionary dependencies. Finally, a binary target variable was added to each example, 1 for a methylated and 0 for a non-methylated site. Two matrices were constructed: M for the positive (methylated) sites and NM for the negative (non-methylated) sites.

Before predictor construction, a t-test feature selection algorithm was applied on each individual feature. Then, after the z-score normalization, we applied the principal component analysis to further reduce dimensionality of the sample. The principal component analysis eliminates correlated features. The preprocessed data was then fed into a support vector machine software *SVM^{light}* [38].

D. Predictor Evaluation

The method used for the accuracy estimation was leave-one-out, which uses all proteins except one to train, while the one left out is used to test. This is performed for each protein, so a different one is left out each time, and each of its positive and negative sites are used for accuracy estimation.

We measured the sensitivity (*sn*) and specificity (*sp*) for the parameters used for the predictor to evaluate its performance. Sensitivity is defined as the percentage of positive (methylated) examples correctly predicted, and specificity is the percentage of negative (non-methylated) examples correctly predicted. The accuracy is determined by the arithmetic mean of sensitivity and specificity and is not affected by the class imbalance. Sensitivity, specificity, and accuracy were measured both per protein (average for all sites in a specific protein) and per residue. In the latter case, each protein will influence the test statistics commensurate with its length, more precisely, to the numbers of arginines and lysines. In addition to accuracy, we also report on the area under the receiver operating characteristic (ROC). The ROC curve is a plot of sensitivity vs. (1 – specificity) and was generated by shifting the decision threshold. The area under the ROC curve (AUC) was estimated using the trapezoid rule.

A. Functional Analysis of Methylated Proteins

In order to determine whether our set of non-redundant methylated proteins was reasonably diverse for analysis and predictor construction, we searched for Gene Ontology (GO) annotations [39]. We considered two proteins to be non-redundant if their pairwise sequence identity was below 30%. The GO annotations are derived from a controlled vocabulary for the following tree categories: biological process, cellular component, and molecular function; and a single protein, or a gene product in general, can have many GO terms associated with it.

For the 33 non-redundant arginine proteins, we observed 47 different categories for biological process, 16 categories for cellular component and 40 categories for molecular function. The most prominent category within the biological process was “mRNA processing”, while five or more proteins were also observed for categories “transcription,” “regulation of transcription, DNA-dependent,” and “nuclear mRNA splicing, via spliceosome.” The most associated cellular component term was “nucleus” (26 proteins), while “ribonucleoprotein complex” and “cytoplasm” contained 7 and 4 proteins, respectively. The most associated term for molecular function was “RNA binding,” with “nucleic acid binding” and “nucleotide binding” following. Three or more proteins contained the following annotations: “DNA binding,” “protein binding,” “transcription coactivator activity,” and “metal ion binding.”

For the 34 non-redundant lysine proteins, we observed 19 different categories for biological process, 20 categories for cellular component and 29 categories for molecular function. The largest represented term for the biological processes was “protein biosynthesis”, followed by “electron transport” and “transport.” For cellular components, the most associated terms were “ribosome,” “ribonucleoprotein complex,” and “intracellular.” The most associated term for molecular function was “structural constituent of ribosome,” with “oxidoreductase activity” and “actin binding” close behind. In summary, these data indicate that a set of proteins studied herein is moderately diverse across all three GO classifications and thus suitable for analysis and predictor construction.

B. Structural Analysis of Methylated Proteins

In order to learn about structural preferences of experimentally verified methylation sites, we performed sequence alignments between proteins in our dataset and all proteins from the Protein Data Bank (PDB) [40]. We looked for such hits where sequence identity was at least 70% in order to provide reasonable accuracy of functional transfer by sequence similarity [41]. Those proteins from PDB that met this criterion were analyzed to see whether the residues experimentally determined to be methylated were covered by the hits from PDB. If so, we looked to see whether structural information (α -helix, β -sheet or coil) existed for these residues. However, we trusted the structural information only for the residues whose flanking regions, five residues upstream and downstream from the methylation site, were not in crystal contacts [42]. Finally, we searched for the experimental evidence of intrinsic disorder, either from missing residues in the PDB file or similarity to proteins from the DisProt database [2].

TABLE I.I-II

HITS BETWEEN METHYLATED PROTEINS FROM SWISS-PROT AND PDB, FOR ARGININE (TABLE I.I) AND LYSINE (TABLE I.II). THE COLUMNS ARE: SWISS-PROT ID, THE PDB ID OF THE BEST HIT, THE COVERAGE OF THE ORIGINAL PROTEIN, THE CORRESPONDING COORDINATES OF THE PDB HIT, THE PERCENTAGE OF SEQUENCE IDENTITY, THE METHYLATED RESIDUES IN THE ORIGINAL PROTEIN AND THOSE FOUND TO BE COVERED IN THE PDB HIT. SECTION COMMENTS DISCUSSES RESIDUES FOUND TO BE IN CRYSTAL CONTACTS, INTERCHAIN CONTACTS AND THOSE WITH EXPERIMENTALLY DETERMINED DISORDER.

SWISS-PROT ID	PDB ID	Coverage by PDB chain	PDB coordinates	Sequence Identity (%)	Methylated Residue	Residue from PDB	Comments
EP300_HUMAN	1SB0 A	566-652	1-87	89	R580	R15, helix	stabilized by interchain contacts
				89	R604	R39, helix	-
PABP4_HUMAN	1G9L	510-643	6-139	78	R518	R14, coil	-

SWISS-PROT ID	PDB ID	Coverage by PDB chain	PDB coordinates	Sequence Identity (%)	Methylated Residue	Residue from PDB	Comments
CALM_EUGGR	1QTX A	1-148	1-148	90	K115, K148	K115, K148, coils	K115, K148 in crystal contacts
CALM_PARTE	1N0Y B	1-148	1-148	100	K13, K115	K13, disordered; K115, coil	K115 in crystal contacts
CAVP_BRALA	1C7W A	81-161	1-81	100	K95, K116	K15, K35, helices	K15, K35 in crystal contacts
CISY_PIG	4CTS B	28-464	1-437	100	K395	K368, coil	-
CYC1_YEAST	2PCC D	1-108	1-108	100	K77	K72, helix	entirely disordered [2]
CYC_ABUTH	1CCR	1-111	2-112	87	K80, K94	K8, K95, helices	K95 in crystal contacts; disordered
DN72_SULAC	1WD1 A	2-64	3-65	93	K6	K7, sheet	K7 in crystal contacts
DN72_SULSO	1R83 A	1-62	1-62	96	K4, K6, K60, K62	K4, K6, K60, K62, coils	all in crystal contacts
	1BNZ A	1-63	2-64	100	K63	K63, coil	K63 in crystal contacts
EF1A1_RABIT	1G7C A	1-443	1-441	81	K36, K55, K79, K318	K36 helix, K55 coil, K79 disorder, K316 helix	K30, K41, K55 and surrounding residues in crystal contacts
FER1_SULTO	1XER	1-103	1-103	99	K29	X29, coil	V25-P27, V30 in crystal contacts
H32_MEDSA	1EQZ G	1-135	2-136	96	K4, K9, K14, K18, K23, K27	K18, K23, K27, coils	K23, K27 in crystal contacts; M0-A15 disordered
H3_CAEEL	1EQZ G	1-135	2-136	97	K27, K36, K79	K27, K36, K79, coils	K27, V35, V77, T80 in crystal contacts
H4_OLILU	1P3F F	17-102	17-102	96	K79	K79, coil	-
H4_PSAMI	1EQZ H	17-102	18-103	98	K20	K20, coil	-
MYH11_CHICK	1BR2 F	1-791	1-791	99	K127	K127, coil	-
MYSS_CHICK	2MYS A	1-843	1-843	93	K35, K130, K551	X35, X130, coils; X551, helix	-
PRO1A_ACACA	2PRF	1-125	1-125	100	K103	K103, coil	-
RBL_TOBAC	4RUB D	1-477	1-477	100	K14	K14, coil	-
RL7_DESVM	1RQS A	53-126	1-74	75	K76, K87	K24, K35, helices	-

Table I summarizes structural information on the non-redundant subset of methylated proteins. In the case of arginine, only two protein hits could be found that cover methylated residues, and of the two sites with reliable structure one was observed in a helical region, while the other was in the loop. In the case of lysine, out of 39 residues covered by the PDB hits, 11 had reliable structure. Of those, 3 were observed in helical regions, while 8 were observed in the loops. Thus, even though little structural information about protein methylation is available from the structural data, it seems that there is a preference for the coil regions.

C. Statistical Analysis of Flanking Regions

We analyzed position-specific amino acid preferences around methylated and non-methylated sites. Standard t-test was used to characterize amino acid enrichments and depletions. An enriched residue at a specific position occurs more frequently in the positive dataset than in the negative, and vice versa for a

depleted residue. Table II shows the top ten enriched and depleted amino acids for arginine, determined by the lowest p-values (cutoff 0.05). We found 70 significant differences between the positive and negative sets with p-values less than 0.05. A distinctive mark of arginine methylation sites is distribution of glycines around the modification site. For the lysine data, we found 38 statistically significant differences, however the p-values were higher (statistical significance is thus lower). Table III lists the top 10 enriched and depleted sites surrounding the lysine. Interestingly, enrichment of glycines is observed upstream from this modification site as well.

In addition to position-specific amino acid preferences, we also analyzed features with high discriminatory power between methylated and non-methylated residues. The top 10 properties are summarized in Tables IV-V for arginine and lysine residues, respectively.

Arginine methylation is characterized by higher scores of disorder predictions (VL2 and VL3 predictors). Of the three

flavors under the VL2 model, VL2-C was selected as the most discriminating one. In addition, high net charge and enrichment in glycines in the close neighborhoods is another fingerprint of methylated arginine sites. Interestingly, glycine residues are typically better conserved near methylated, as compared to non-methylated sites. Negatively correlated features are represented by sequence complexity measures and conservation of hydrophobics leucine and isoleucine, and charged residues glutamic acid and lysine.

TABLE II
TOP 10 AMINO ACIDS ENRICHED AND DEPLETED AROUND KNOWN ARGININE METHYLATION SITES AS DETERMINED BY THE T-TEST.

Enriched			Depleted		
Position	Residue	p-value	Position	Residue	p-value
+1	G	$2.5 \cdot 10^{-66}$	+4	E	$8.7 \cdot 10^{-3}$
+2	G	$2.8 \cdot 10^{-14}$	-11	E	$1.1 \cdot 10^{-2}$
+4	G	$9.8 \cdot 10^{-10}$	+1	Q	$1.2 \cdot 10^{-2}$
-2	G	$1.1 \cdot 10^{-9}$	+3	E	$1.3 \cdot 10^{-2}$
-4	G	$1.1 \cdot 10^{-9}$	+5	E	$1.3 \cdot 10^{-2}$
-1	G	$1.9 \cdot 10^{-9}$	+1	E	$1.4 \cdot 10^{-2}$
+6	G	$3.5 \cdot 10^{-9}$	-4	E	$1.7 \cdot 10^{-2}$
+11	G	$5.7 \cdot 10^{-7}$	-8	Q	$1.7 \cdot 10^{-2}$
+10	G	$6.3 \cdot 10^{-7}$	+9	E	$2.0 \cdot 10^{-2}$
+7	G	$7.3 \cdot 10^{-7}$	-5	E	$2.1 \cdot 10^{-2}$

TABLE III
TOP 10 AMINO ACIDS ENRICHED AND DEPLETED AROUND KNOWN LYSINE METHYLATION SITES AS DETERMINED BY THE T-TEST.

Enriched			Depleted		
Position	Residue	p-value	Position	Residue	p-value
-2	P	$4.1 \cdot 10^{-5}$	+2	K	$2.5 \cdot 10^{-2}$
-1	F	$9.2 \cdot 10^{-5}$	-8	L	$3.0 \cdot 10^{-2}$
-7	T	$7.8 \cdot 10^{-4}$	-4	A	$3.5 \cdot 10^{-2}$
-2	G	$9.2 \cdot 10^{-4}$			
-11	G	$1.1 \cdot 10^{-3}$			
+12	F	$1.6 \cdot 10^{-3}$			
-10	P	$2.7 \cdot 10^{-3}$			
-5	I	$3.5 \cdot 10^{-3}$			
+5	Y	$3.6 \cdot 10^{-3}$			
-8	T	$5.4 \cdot 10^{-3}$			

TABLE IV
TOP 10 FEATURES, POSITIVELY AND NEGATIVELY CORRELATED WITH ARGININE METHYLATION. NUMBERS IN PARENTHESES INDICATE WINDOW SIZES. SEE TEXT FOR DETAILED EXPLANATIONS.

Positive correlation		Negative correlation	
Feature	p-value	Feature	p-value
residue G (3)	$7.4 \cdot 10^{-25}$	PSSM, K (1)	$2.1 \cdot 10^{-9}$
PSSM, G (5)	$1.5 \cdot 10^{-24}$	PSSM, E (11)	$3.2 \cdot 10^{-9}$
B-factor (1)	$5.8 \cdot 10^{-8}$	β -entropy (25)	$5.8 \cdot 10^{-9}$
VL3 (11)	$5.8 \cdot 10^{-8}$	entropy (25)	$6.0 \cdot 10^{-9}$
VL2-C (1)	$8.5 \cdot 10^{-8}$	charge (3)	$1.2 \cdot 10^{-8}$
net charge (25)	$2.0 \cdot 10^{-7}$	residue L (25)	$4.0 \cdot 10^{-6}$
PSSM, S (11)	$2.7 \cdot 10^{-6}$	residue E (25)	$4.8 \cdot 10^{-6}$
hydro. moment (11)	$2.9 \cdot 10^{-6}$	residue D (25)	$8.9 \cdot 10^{-6}$
PSSM, P (11)	$2.6 \cdot 10^{-5}$	PSSM, I (11)	$2.7 \cdot 10^{-5}$
PSSM, R (11)	$4.1 \cdot 10^{-4}$	PSSM, L (5)	$5.1 \cdot 10^{-5}$

Lysine methylation is characterized by a different set of top-ranked features than arginine sites, except for the high negative charge, high evolutionary conservation of glycine, and conservation of proline. On the other hand, non-methylated lysine sites are characterized by a low conservation of

several residues. Note that the p-values for lysine methylation are orders of magnitude higher than those for arginines, indicating much higher confidence in discriminatory attributes. This is likely a consequence of noisier data and/or a more difficult classification problem.

TABLE V
TOP 10 FEATURES, POSITIVELY AND NEGATIVELY CORRELATED WITH LYSINE METHYLATION. NUMBERS IN PARENTHESES INDICATE WINDOW SIZES. SEE TEXT FOR DETAILED EXPLANATIONS.

Positive correlation		Negative correlation	
Feature	p-value	Feature	p-value
PSSM, P (11)	$6.8 \cdot 10^{-8}$	PSSM, L (11)	$1.3 \cdot 10^{-6}$
residue P (7)	$8.3 \cdot 10^{-6}$	PSSM, E (11)	$3.6 \cdot 10^{-6}$
residue Y (19)	$9.0 \cdot 10^{-6}$	PSSM, Q (11)	$2.6 \cdot 10^{-5}$
PSSM inf/pos (5)	$1.0 \cdot 10^{-5}$	residue E (25)	$5.5 \cdot 10^{-5}$
aromaticity (3)	$2.5 \cdot 10^{-5}$	VL2-S (11)	$5.8 \cdot 10^{-5}$
residue F (3)	$3.2 \cdot 10^{-5}$	residue L (25)	$1.2 \cdot 10^{-4}$
residue T (19)	$7.5 \cdot 10^{-4}$	charge (7)	$3.2 \cdot 10^{-4}$
residue G (25)	$9.1 \cdot 10^{-4}$	PSSM, M (5)	$9.2 \cdot 10^{-4}$
net charge (25)	$1.8 \cdot 10^{-3}$	PSSM, I (1)	$1.6 \cdot 10^{-3}$
PSSM, G (11)	$2.3 \cdot 10^{-3}$	PSSM, A (11)	$2.2 \cdot 10^{-3}$

D. Methylation and Intrinsic Disorder

We used computational approaches to associate methylation sites with the presence or absence of intrinsic disorder. The average scores of disorder predictions by VL3 model around known methylation sites were thus compared with average disordered predictions at all other arginines/lysines in the same set of proteins. The results of these comparisons are shown in Figure 1. Interestingly, arginine residues are strongly predicted to be in disordered regions, having an average score of 0.83 ± 0.02 vs. an average score of 0.66 ± 0.01 for the remaining arginines. On the other hand, the respective scores in the lysine case were 0.55 ± 0.04 vs. 0.57 ± 0.01 . We applied the same predictor to the long, experimentally verified, disordered and ordered regions and obtained scores 0.75 ± 0.01 and 0.21 ± 0.01 . These results indicate that both arginine and lysine methylation of currently verified sites are highly likely to prefer structural disorder. However, it appears that the proteins from the lysine dataset contain a significant fraction of disorder even for non-methylated lysines, which makes them hard to distinguish based on this property alone. For example, myosin contains several experimentally confirmed disordered regions, while apocytochrome C (cytochrome C with the heme removed) is known to be completely disordered [43, 44]. Both proteins are present in our lysine set.

E. Predictor Evaluation

Extensive experiments were performed in predictor construction. In Tables VI-VII we show prediction results for arginine and lysine methylation, respectively. These estimates were obtained using a per-protein leave-one-out methodology. In each table, the thresholds for feature selection filter were set to 0.1, while the number of principal components was kept at 20 in order to eliminate many correlated features and significantly reduce dimensionality of the sample. We evaluated only polynomial kernels, while the default value was used for the learning parameter C [38]. The improvement in prediction accuracy provided by using features related to protein flexibility, and adding feature selection and principal

component analysis was approximately 4 percentage points both for arginine and lysine datasets.

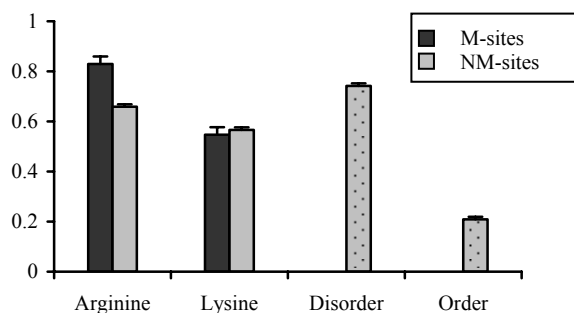


Figure 1. Average scores of VL3 disorder predictor on methylated (dark grey bars) and non-methylated (light grey bars) residues. The dotted single bars under Disorder and Order show VL3 predictions on experimentally confirmed long disordered regions and long ordered regions, respectively. VL3 outputs scores between zero and one. Error bars represent 68% confidence intervals.

It can be easily observed from Tables VI-VII that the prediction accuracy achieved by the various models generally drops with the increase of the polynomial degree. We have also observed (data not shown) that the accuracy continued to decrease when the degree of the polynomial kernel was further incremented, up to 5. This behavior was expected, since the datasets were small and the number of features relatively high (263 in total, 20 after the principal component analysis). Further increases in dataset sizes may likely enable more accurate learning using polynomials of higher degree.

TABLE VI
CLASSIFICATION ACCURACY [%] FOR THE PREDICTION OF ARGININE METHYLATION SITES; *sn* – SENSITIVITY, *sp* – SPECIFICITY, *acc* = (*sn* + *sp*) / 2 – ACCURACY, *AUC* – AREA UNDER THE ROC CURVE.

<i>polynomial degree</i>	<i>Arginine methylation</i>			<i>AUC</i>
	<i>sn</i>	<i>sp</i>	<i>acc</i>	
<i>p</i> = 1	79.0	76.7	77.8	83.9
<i>p</i> = 2	73.6	82.2	77.9	85.0
<i>p</i> = 3	47.9	86.8	67.3	73.3

TABLE VII
CLASSIFICATION ACCURACY [%] FOR THE PREDICTION OF LYSINE METHYLATION SITES; *sn* – SENSITIVITY, *sp* – SPECIFICITY, *acc* = (*sn* + *sp*) / 2 – ACCURACY, *AUC* – AREA UNDER THE ROC CURVE.

<i>polynomial degree</i>	<i>Lysine methylation</i>			<i>AUC</i>
	<i>sn</i>	<i>sp</i>	<i>acc</i>	
<i>p</i> = 1	65.9	60.4	63.1	66.4
<i>p</i> = 2	33.6	73.4	53.5	58.4
<i>p</i> = 3	35.6	86.3	61.0	68.3

IV. DISCUSSION

In this paper we have analyzed reversible protein methylation and provided evidence that it is predictable from amino acid sequence. This is useful since computational approaches combined with good biological insight can be important in not only detecting proteins that are likely to be methylated, but also finding the exact sites of modification. Our results indicate that arginine methylation sites are predictable with significantly higher accuracy than the lysine sites. The most likely

reason for such performance is an interplay of the quality of the dataset and our assumptions that all sites not labeled as positive were consequently negative. In the case of lysine residues, this assumption is less likely to be correct since only individual experiments were performed on each protein. On the other hand, proteomics studies were used to detect arginine methylation which left much less room for noise. In addition, of course, it is also possible that prediction of lysine methylation is simply a more difficult problem than prediction of arginine methylation due to greater similarities in the sequences around methylated and non-methylated lysines.

Clearly, in order to provide a better automated identification of methylation sites, more experimentally verified sites are needed. In addition, the emergence of proteomics data will likely lead to significant improvement in our understanding of sequence preferences around methylation sites, especially for lysine methylation. This would also help to diversify current datasets, since a large number of the examples in our original dataset are coming from four non-redundant homologous groups: calmodulin, cytochrome C, rubisco large subunit, and histones. Also, determining which methyltransferases methylate each residue might prove useful since it is hypothesized that their activity is highly specific. Some sites in the SWISS-PROT database have annotations specifying which methyltransferase added the methyl group, but not for all proteins and not in a standard format. A method of extracting that data from SWISS-PROT as well as mining literature would be necessary.

Our working hypothesis is that protein modification sites prefer flexible and intrinsically disordered protein regions. In our prior work, we already strongly associated phosphorylation sites with such regions [45], especially in the cases of serine and threonine phosphorylation. It has been hypothesized that about 1/3 of all eukaryotic proteins undergoes protein phosphorylation [46] so identifying the actual sites of modification is of great importance. We used computational approaches to estimate that these fractions are significantly higher for several classes of proteins, e.g. those classes associated with regulation, transcription, or even cancer. The improvements in computational approaches will soon enable such estimates regarding protein methylation. In the case of methylation sites, we provide evidence that both arginine and lysine methylation prefer structurally flexible regions and intrinsic disorder rather than order, at least for the currently identified sites. This evidence is partially based on anecdotal observations (e.g. p53, cytochrome C, or hnRNP core protein A1 – see Table I.II), but is predominantly of statistical nature. However, both predictors used in this study are characterized by a relatively small error rate, especially on long ordered and disordered regions [35], and thus are unlikely to fail big.

Our working hypothesis regarding the importance of disorder for protein modification arose from empirical evidence that many structurally characterized regions of disorder contained sites of modification [24]. Reflection on this observation, on the results herein, and on the results of our previous work on phosphorylation [45] suggest a few possible advantages for locating sites of modification in disordered regions. First, even exactly the same local amino acid sequence in two non-homologous structured proteins typically adopts different secondary structures [47] and so would present different

shapes to the modifying enzyme. In such a circumstance, the same sequence from the different proteins would not fit into one enzyme active site. On the other hand, if similar (not to mention identical) local sequences were to exist in regions of disorder, each could change its conformation to fit into the active site of a single enzyme. Second, the surfaces of structured proteins are typically smooth and nearby residues are often buried into the folded core. In this second circumstance, it is again difficult to understand specific modification. On the other hand, if the residue to be modified were within a disordered region, the exposed surface area would be very large and convoluted in a sequence-dependent manner, a situation that would facilitate highly specific, sequence-dependent association with the modifying enzyme. Finally, the chemical modification would typically cause only small changes on the surface of a structured protein. On the other hand, the chemical modification could cause large-scale changes in a disordered region, such as the induction of a disorder-to-order transition. A large change in structure would be a less ambiguous signal as compared to a small change.

Many protein modifications can work together in regulation and signaling, so we expect that combining predictors for methylation, acetylation, and phosphorylation, as well as other protein modifications would be very useful in future. As an example of such a process, it has been shown that the phosphorylation of serine 10 of the H3 histone protein prevents the methylation of lysine 9 [48]. This will be one focus of our future work.

REFERENCES

- [1] Paik, W.K. and S. Kim, *Enzymatic methylation of protein pabp1 identified as an arginine fractions from calf thymus nuclei*. *Biochem. Biophys. Res. Commun.*, 1967. 29: p. 14-20.
- [2] Vucetic, S., et al., *DisProt: a database of protein disorder*. *Bioinformatics*, 2005. 21(1): p. 137-140.
- [3] Paik, W.K. and S. Kim, *Enzymology of protein methylation*. *Yonsei Medical Journal*, 1986. 25(3): p. 159-177.
- [4] Aletta, J.M., T.R. Cimato, and M.J. Ettinger, *Protein methylation: a signal event in post-translational modification*. *Trends Biochem. Sciences*, 1998. 23(3): p. 89-91.
- [5] Najbauer, J. and D. Aswad, *Diversity of methyl acceptor proteins in rat pheochromocytoma (PC12) cells revealed after treatment with adenosine dialdehyde*. *J. Biol. Chem.*, 1990. 265: p. 12717-12721.
- [6] Shi, Y., et al., *Histone demethylation mediated by the nuclear amine oxidase homolog LSD1*. *Cell*, 2004. 119: p. 941-953.
- [7] Bedford, M.T. and S. Richard, *Arginine methylation: an emerging regulator of protein function*. *Mol. Cell*, 2005. 18: p. 263-272.
- [8] Scorilas, A., et al., *Genomic organization, physical mapping, and expression analysis of the human protein arginine methyltransferase 1 gene*. *Biochem. Biophys. Res. Commun.*, 2000. 278: p. 349-359.
- [9] Najbauer, J., et al., *Peptides with sequences similar to glycine, arginine-rich motifs in proteins interacting with rna are efficiently recognized by methyltransferase(s) modifying arginine in numerous proteins*. *J. Biol. Chem.*, 1993. 268: p. 10501-10509.
- [10] Lachner, M. and T. Jenuwein, *The many faces of histone lysine methylation*. *Curr. Opin. Cell. Bio.*, 2002. 14(3): p. 286-298.
- [11] Friesen, W.J., et al., *The methylosome, a 20S complex containing JBP1 and pICln, produces dimethylarginine-modified Sm proteins*. *Mol. Cell. Biol.*, 2001. 21: p. 8289-8300.
- [12] Cheng, D., et al., *Small molecule regulators of protein arginine methyltransferases*. *J. Biol. Chem.*, 2004. 279: p. 23892-23899.
- [13] Bairoch, A., et al., *The Universal Protein Resource (UniProt)*. *Nucleic Acids Res*, 2005. 33 Database Issue: p. D154-9.
- [14] Murray, K., *The occurrence of var epsilon-n-methyl lysine in histones*. *Biochemistry*, 1964. 3: p. 10-15.
- [15] Strahl, B.D. and C.D. Allis, *The language of covalent histone modifications*. *Nature*, 2000. 403(6765): p. 41-45.
- [16] van Holde, K., *Chromatin*. 1988, New York: Springer-Verlag.
- [17] Chuikov, S., et al., *Regulation of p53 activity through lysine methylation*. *Nature*, 2004. 432: p. 353-360.
- [18] Dyson, H.J. and P.E. Wright, *Intrinsically unstructured proteins and their functions*. *Nat Rev Mol Cell Biol*, 2005. 6(3): p. 197-208.
- [19] Fink, A.L., *Natively unfolded proteins*. *Curr Opin Struct Biol*, 2005. 15(1): p. 35-41.
- [20] Dunker, A.K., et al., *Intrinsically disordered protein*. *J. Mol. Graph. Model.*, 2001. 19(1): p. 26-59.
- [21] Dedmon, M.M., et al., *FlgM gains structure in living cells*. *Proc. Natl. Acad. Sci. U. S. A.*, 2002. 99(20): p. 12681-12684.
- [22] Pielack, G., *personal communication*. 2005.
- [23] Dunker, A.K. and Z. Obradovic, *The protein trinity - linking function and disorder*. *Nat. Biotechnol.*, 2001. 19(9): p. 805-806.
- [24] Dunker, A.K., et al., *Intrinsic disorder and protein function*. *Biochemistry*, 2002. 41(21): p. 6573-6582.
- [25] Tompa, P., *Intrinsically unstructured proteins*. *Trends Biochem Sci*, 2002. 27(527-533).
- [26] Tompa, P., *personal communication*. 2004.
- [27] Daughdrill, G.W., et al., *Natively disordered protein*, in *Protein Folding Handbook*, J. Buchner and T. Kiefhaber, Editors. 2005, Wiley-VCH: Verlag GmbH & Co. KGaA: Weinheim. p. 271-353.
- [28] Plewczynski, D., et al., *AutoMotif server: prediction of single residue post-translational modifications in proteins*. *Bioinformatics*, 2005. 21(10): p. 2525-7.
- [29] Ong, S.-E., G. Mittler, and M. Mann, *Identifying and quantifying in vivo methylation sites by heavy methyl silac*. *Nat. Methods*, 2004. 1: p. 119-126.
- [30] Vihinen, M., E. Torkkila, and P. Riikonen, *Accuracy of protein flexibility predictions*. *Proteins*, 1994. 19: p. 141-149.
- [31] Eisenberg, D., R.M. Weiss, and T.C. Terwilliger, *The hydrophobic moment detects periodicity in protein hydrophobicity*. *Proc. Natl. Acad. Sci. U. S. A.*, 1984. 81: p. 140-144.
- [32] Wootton, J.C. and S. Federhen, *Analysis of compositionally biased regions in sequence databases*. *Methods Enzymol*, 1996. 266: p. 554-571.
- [33] Daroczy, Z., *Generalized information functions*. *Information and Control*, 1970. 16: p. 36-51.
- [34] Vucetic, S., et al., *Flavors of protein disorder*. *Proteins*, 2003. 52: p. 573-584.
- [35] Obradovic, Z., et al., *Predicting intrinsic disorder from amino acid sequence*. *Proteins*, 2003. 53(S6): p. 566-572.
- [36] Radivojac, P., et al., *Protein flexibility and intrinsic disorder*. *Protein Science*, 2004. 13(1): p. 71-80.
- [37] Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs*. *Nucleic Acids Res.* 1997. 25: p. 3389-3402.
- [38] Joachims, T., *Learning to classify text using support vector machines: methods, theory, and algorithms*. 2002: Kluwer Academic Publishers.
- [39] Ashburner, M., et al., *Gene ontology: tool for the unification of biology*. *The Gene Ontology Consortium*. *Nat. Genet.*, 2000. 25(1): p. 25-29.
- [40] Berman, H.M., et al., *The protein data bank*. *Nucleic Acids Res.*, 2000. 28(1): p. 235-242.
- [41] Rost, B., et al., *Automatic prediction of protein function*. *Cell Mol Life Sci*, 2003. 60(12): p. 2637-2650.
- [42] Sobolev, V., et al., *Automated analysis of interatomic contacts in proteins*. *Bioinformatics*, 1999. 15(4): p. 327-32.
- [43] Babul, J. and E. Stellwagen, *Participation of the protein ligands in the folding of cytochrome c*. *Biochemistry*, 1972. 11(7): p. 1195-200.
- [44] Fisher, W.R., H. Taniuchi, and C.B. Anfinsen, *On the role of heme in the formation of the structure of cytochrome c*. *J Biol Chem*, 1973. 248(9): p. 3188-95.
- [45] Iakoucheva, L.M., et al., *The importance of intrinsic disorder for protein phosphorylation*. *Nucleic Acids Res.*, 2004. 32(3): p. 1037-1049.
- [46] Johnson, L.N. and R.J. Lewis, *Structural basis for control by phosphorylation*. *Chem. Rev.*, 2001. 101: p. 2209-2242.
- [47] Kabsch, W. and C. Sander, *On the use of sequence homologies to predict protein structure: identical pentapeptides can have completely different conformations*. *Proc. Natl. Acad. Sci. U. S. A.*, 1984. 81(4): p. 1075-8.
- [48] Pal, S., et al., *Human SWI/SNF-associated PRMT5 methylates histone H3 arginine 8 and negatively regulates expression of ST7 and NM23 tumor suppressor genes*. *Mol. Cell. Biol.*, 2004. 24: p. 9630-9645.