Endocrine, *vol. 11, no. 3, 285–292, December 1999* 0969–711X/99/11:285–292/$12.00

Reevaluation of the Determinants of Tyrosine Sulfation

Hugh B. Nicholas, Jr.,[1] Steve S. Chan,[2,3] and Grace L. Rosenquist[2]

[1]Pittsburgh Supercomputing Center, Pittsburgh, PA; [2]Section of Neurobiology, Physiology and Behavior, University of California, Davis, CA; and [3]Hahnemann University School of Medicine, Philadelphia, PA

The posttranslational sulfation of tyrosine has been thought to be initiated by the recognition of specific consensus features by the sulfating enzyme tyrosyl-protein sulfotransferase (TPST). However, using these recognition features to identify new tyrosine sulfation sites misses recently characterized sites that lack these features. Rigorous analysis of the amino acids surrounding the target tyrosine using the position-specific scoring matrix (PSSM) demonstrates that a consensus sequence does not contain all the information necessary to predict tyrosine sulfation. Instead, accurate prediction requires consideration of all residues within five amino acids on either side of the target tyrosine. These results support the notion that secondary structure is the major determinant of sulfation and that other residues within the sulfation site can compensate for deviations from commonly observed features. This view implies that specific consensus features are not critical for TPST substrate recognition but that TPST may instead broadly recognize any sufficiently exposed tyrosine residue.

Key Words: Tyrosine sulfation; posttranslational processing; tyrosyl protein sulfotransferase.

Introduction

Tyrosine sulfation is a common posttranslational modification of proteins transported through the Golgi system. Sulfation regulates the function of a number of important proteins including gastrin , C4 complement, and coagulation factor VIII. The most common effect of tyrosine sulfation is to enhance protein activity. For example, sulfation increases proteolysis of the various gastrin forms *(1)*, increases the hemolytic activity of C4 complement *(2)*, and increases the procoagulant activity of Factor VIII *(3)*. Yet, despite estimates that up to 1% of all eukaryotic proteins may be sulfated *(4)*, tyrosine sulfation has been experimentally verified in only about 100 of the roughly 200,000 tyrosine sites in the Swiss-Prot database. Of these, only gastrin and cholecystokinin among the mammalian regulatory proteins, which include hormones and cytokines, are known to be sulfated. Over the years, a consensus sequence based on empirical observations of sulfation sites has been developed in an attempt to identify the remaining sulfated tyrosines *(5,6)*. Essentially, the consensus sequence specified allowed and disallowed amino acids near sulfated tyrosines. Yet, as the number of confirmed instances of sulfated tyrosines has grown, so has the number of sulfation sites that do not conform to established consensus features. Indeed, although specific types of amino acids are highly conserved, complete conservation exists at no position within the sulfation site.

Our current work replaces the empirical consensus sequence with a position-specific-scoring matrix (PSSM) representation of the tyrosine sulfation site. The PSSM is able to incorporate observed exceptions into a unified analytical framework in which they can be systematically examined. Advances in statistical and mathematical bioinformatics *(7)* allowed our PSSM to efficiently capture information from specific protein sequences containing known sulfation sites and contrast this with information from sites in which the tyrosine is not sulfated. We use this PSSM to predict 18 sulfation sites in 16 hormones, neuropeptides, and cytokines.

This reevaluation of the determinants of tyrosine sulfation sites addresses the growing number of experimentally confirmed sulfation sites that do not follow conventional rules governing tyrosine sulfation. Our findings suggest that the tyrosyl-protein sulfotransferase (TPST) recognition of tyrosines is broadly selective and will recognize any suitably accessible tyrosine. Finally, we present the PSSM as an efficient statistical description of the differences between two classes of substrates, tyrosine sites that undergo sulfation and sites that do not. Ultimately, our goal is to understand the molecular basis of tyrosine sulfation by developing techniques for predicting the occurrence of tyrosine sulfation in proteins and peptides.

Received September 7, 1999; Revised October 4, 1999; Accepted October 4, 1999.

Author to whom all correspondence and reprint requests should be addressed: Dr. Grace L. Rosenquist, Section of Neurobiology, Physiology and Behavior, University of California, One Shields Avenue, Davis, CA 95616.

E-mail: rosenqui@a.psc.edu

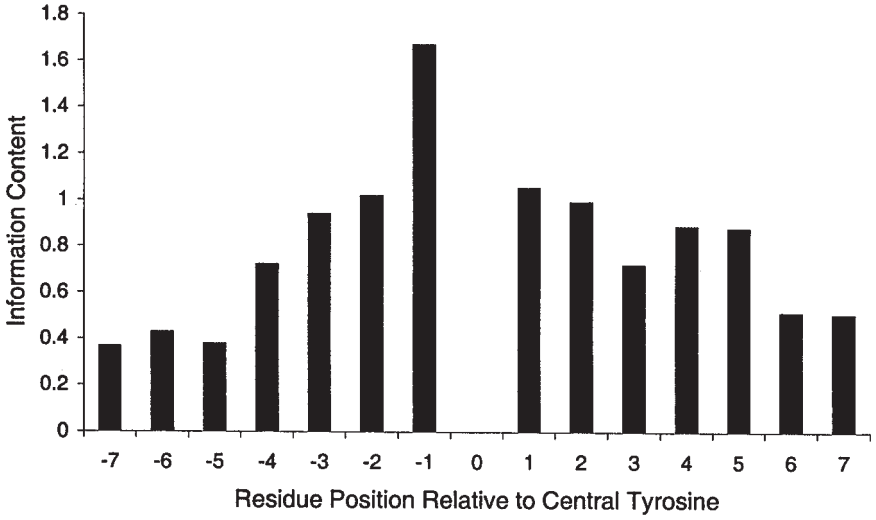


Fig. 1. Information content. The information content at each position was computed by the method of Stormo and Hartzell. The highest value occurs at position –1 relative to the tyrosine and represents the position with the highest information content. The information content measures the differences in the kinds of amino acids present at a specific position relative to tyrosines that can be enzymatically sulfated and those that cannot be enzymatically sulfated. A higher information content value indicates a greater difference.

Results

The information content (relative entropy) was highest at the –1 position relative to tyrosine. Previous observations by others and by us have indicated that acidic residues at this position predominate in sulfated tyrosine sites *(5,7,8)* (Fig. 1). The information content plot indicates that the highest information content occurs in a window from position –4 to position +5. Because of the convenience of using a symmetrical window, we will use the window from –5 to +5 rather than in the –7 to +7 window previously used.

The Pattern/Negative (M = 13) PSSM was used to examine and score 189,195 tyrosine sites from the Swiss-Prot Metazoa database. Of 103 known tyrosine sulfation sites, 102 scored higher than 2.89. The average score for positive sites (p-sites) was 5.9 ± 4.8 SD. In contrast, negative site (n-site) and Swiss-Prot site (sw-site) scores had averages of –9.50 ± 5.7 SD and –8.6 ± 5.9 SD, respectively (Table 1). A receiver operating characteristic (ROC) plot of the p-sites with the jackknifed PSSM showed that the ROC statistic was 0.9876, close to a perfect discrimination score of 1 (Fig. 2). Thirty jackknifed p-site scores were at least 2.89. This score was set as the minimum threshold for potential sulfation sites. Of the 5594 sw-sites that scored over 2.89, 267 sites were conserved in at least two animals and were known to pass through the endoplasmic reticulum (ER). Passage through the ER is a necessary requirement because tyrosine sulfation occurs in the distal Golgi. Eighteen of these predicted sulfation sites were on 16 peptide hormones, neuropeptides, or cytokines (Table 2); eight of these 16 sulfation sites have an acidic residue at –1.

Cumulative distribution functions (CDFs) were used to visually summarize the distribution of PSSM-generated scores for P, random (rn), sw, and n-sites. Jackknifed scores were generated by scoring the omitted p-site sequence with its jackknifed PSSM. The CDFs of n- and sw-site scores were nearly identical (Fig. 3), consistent with the fact that the vast majority of tyrosines in sw-sites are not sulfated.

Table 1
Scores for p-Sites Using PSSM Constructed with All p-sites Represented

	n-Sites	sw-Sites	rn-Sites	p-Sites (jackknifed)[a]	p-Sites
Average	-9.4	-8.6	3.2	5.9	10.4
SD	5.7	6.0	5.4	4.8	3.9
Low score	-31.0	-36.9	-13.7	-4.3	2.9
High score	7.5	20.0	18.1	14.4	17.7

[a]Jackknifed scoring matrices used for these p-site scores.

The distribution of jackknifed scores and the scores for rn-sites differed enough to be statistically significant (D = 0.230, significance probability = 0.01). The difference in jackknifed scores was clearly distinct from the distribution of scores for n-sites (D = 0.860, significance probability = 4×10^{-24}). The rn-site CDF allowed us to partition the difference between the n-site CDF and the jackknifed score CDF into two components. The first, and by far the largest component, was the difference between the n-site CDF and the rn-site CDF. The difference between these curves corresponded to gross differences in amino acid composition between the sequences in p-sites and n-sites. The distance between the rn-site CDF and the CDF for jackknifed scores was attributable to smaller differences in the population of amino acids at specific positions within the p-sites.

The jackknifed score distribution was used as the reference to compare the other curves in the CDF. The distance between the CDF for the p-sites with the regular PSSM and

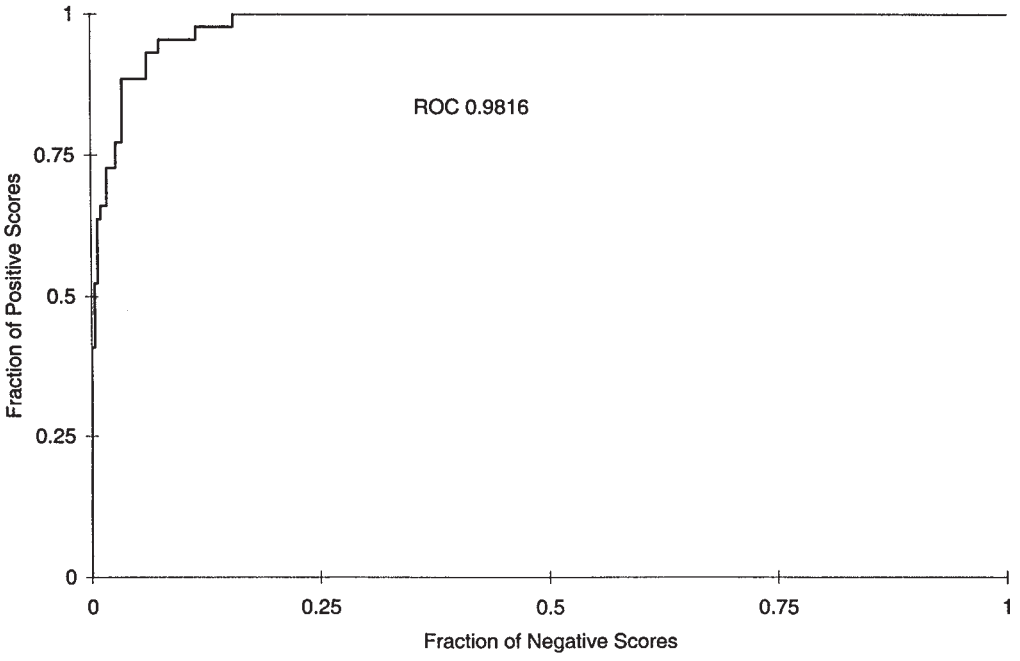


Fig. 2. ROC statistic evaluation of jackknifed PSSM evaluated with the p-sites. Lawrence/Pattern pseudocounts with a multiplier of 13 were added to each position of the PSSM. The ROC statistic (area under the curve) ranges from 0 to 1.0. The ROC statistic measures how effectively the tyrosine sulfation scores generated by our PSSM distinguishes between tyrosines that can be enzymatically sulfated and those that cannot be enzymatically sulfated. A ROC statistic of 1.0 would mean that all of the tyrosines that can be enzymatically sulfated scored higher than any of the tyrosines that cannot be enzymatically sulfated.

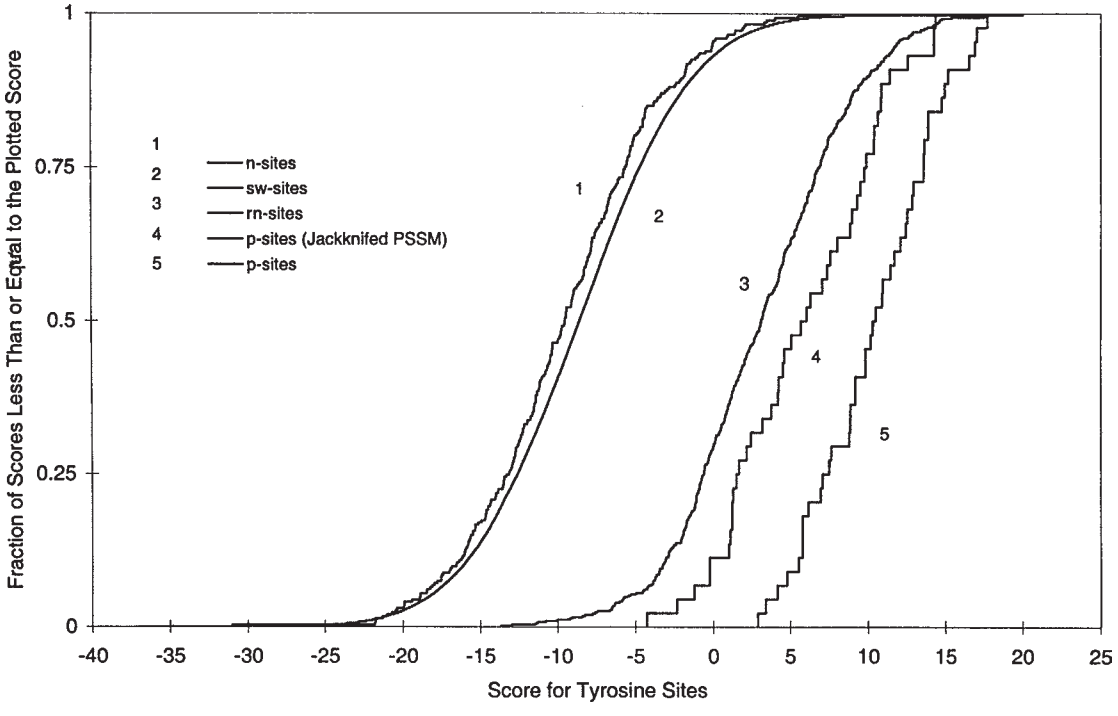


Fig. 3. Cumulative distribution functions were created for the scores for n-sites, sw-sites, rn-sites, and p-sites. Except for one case in which scores from the jackknifed PSSM were used, all scores were generated from the p-site PSSM with Pattern pseudocounts in an n-site background, a window -5 to +5, and a multiplier of 13. Each curve shows the observed range of scores for a specific kind of sequence environment for the central tyrosine. Thus, the difference between curve 1 (n-sites) and curve 4 (p-sites with jackknifed PSSM) reflects the difference that would be expected for tyrosines in a sequence environment where they cannot be enzymatically sulfated (curve 1), and for tyrosines in a sequence environment where they can be enzymatically sulfated (curve 4). The difference between curve 3 (rn-sites) and curve 4 shows how the scores would change if the amino acids in the environment where tyrosines that can be enzymatically sulfated were distributed randomly rather than at the specific positions in which they were found.

Table 2
Scores for Hormones, Neuropeptides, and Cytokines Predicted to Be Tyrosine Sulfated

Common name and function	Swiss-Prot name	Tyrosine sulfation site	Residue number[a]	PSSM score
Fibroblast growth factor-6	FGF6-HUMAN	LpnnynaYesdlyqg	171	3.37
mitogenic and angiogenic functions	FGF6-MOUSE	LpnnynaYesdlyrg	171	3.37
Follistatin	FSA-HUMAN	EededqdYsfpissi	328	5.42
inhibits secretion of FSH	FSA-MOUSE	EeeedqdYsfpissi	334	5.47
Inhibin β A chain	IHBA-HUMAN	rpmsmlyYddgqnii	404	3.59
growth factor	IHBA-RAT	rpmsmlyYddgqnii	402	3.59
Interleukin-1 α	IL1A-MOUSE	cysenedYssaidhl	21	8.26
cytokine: involved in inflammatory response	IL1A-RAT	cyseneeYssaidhl	21	6.42
Interleukin-1 β	IL1B-HUMAN	asemmayYsgneddl	15	5.77
cytokine: involved in inflammatory response	IL1B-BOVIN	inemmayYsdenell	15	3.70
C-Type natriuretic peptide	ANFC-SCYCA	srlledeYghylpsd	19	5.20
vasorelaxant and CGMP-stimulating activity	ANFC-TRISC	srlledeYghylpsd	19	5.20
β-Neoendorphin-dynorphin	NDDB-HUMAN	sqedpnaYsgelfda	247	3.27
neuropeptide	NDDB-RAT	sqenpntYsedldvv	198	3.69
Pancreatic hormone	PAHO-MOUSE	gaplepmYpgdyatp	36	4.06
regulator of pancreatic and gastrointestinal functions	PAHO-RAT	gaplepmYpgdyath	36	4.06
Placental lactogen I	PLC1-MOUSE	eenenfdYpawsgle	172	5.92
mammogenesis and lactogenesis	PLCV-RAT	eenenfdYpawsglk	171	5.92
Proenkephalin A	PENK-HUMAN	mkkmdelYpmepeee	118	4.92
neuropeptide	PENK-MOUSE	mkkmdelYpmepeee	118	4.92
Prolactin[b]	PRL-BUFJA	gdleneiYspwpgps	82	4.42
Lactogenesis	PRL-CHEMY	geienelYspwsglp	146	2.98
Somatolactin	SOML-GADMO	lqttldrYddvpdvl	138	5.73
member of somatotropin/prolactin family	SOML-ONCKE	lqttldrYddapdtl	136	4.72
Somatotropin	SOMA-OREMO	nqdeaenYpdtdtlq	127	8.35
growth hormone	SOMA-ORENI	nqdeaenYpdtdtlq	144	8.35
	SOMA-CTEID	plpfedfYltmgess	166	2.94
	SOMA-CYPCA	plpfedfYltmgenn	166	2.94
Transforming growth factor β2	TGF2-HUMAN	ltsppedYpepeevp	57	10.57
Suppresses interleukin-1 dependent cell growth	TGF2-MOUSE	ltsppedYpepdevp	57	10.63
Tumor necrosis factor α	TNFA-HUMAN	aelnrpdYldfaesg	217	3.82
cytokine	TNFA-PIG	aelnlpdYldfaesg	216	4.32
	TNFA-SHEEP	nlpeyldYaesgqvy	215	3.80
	TNFA-BOVIN	nlpdyldYaesgqvy	220	3.34
Vascular endothelial growth factor	VEGF-HUMAN	lvdifqeYpdeieyi	65	6.17
Angiogenesis and endothelial growth factor	VEGF-MOUSE	lvdifqeYpdeieyi	64	6.17

[a] Residue number refers to the location of the tyrosine residue in the sequence from the Swiss-Prot database.
[b] Sheep prolactin is reported to be tyrosine sulfated (Kohli, R., Chadha, N., Muralidhaar, K. [1988]. *Fed. Eur. Biochem. Soc.* **242,** 139–143).

the jackknifed PSSM reflected the fact that the PSSM was derived from a small amount of data. Each individual sequence still had a relatively large effect on the PSSM, which would become negligible with sufficient data. In the absence of large amounts of data, the jackknifed scores were a more accurate reflection of how the PSSM would perform with a new p-site.

The Kolmogorov-Smirnov test of the score distributions from p-sites and n-sites gave a 10^{-25} significance probability. An additional CDF was generated for the individual sites left out of the jackknifed PSSM. The CDF for the jackknifed scores lay between the p-site CDF and the rn-site CDF. The difference in the average scores between the jackknifed scores and the n-site sequences was 15.3, indicating that the PSSM was able to distinguish between sulfated and nonsulfated tyrosines.

Discussion

Our results demonstrate that sequence information can be used to predict tyrosine sulfation by TPST. We have shown previously that the widely used heuristic rules for identifying sulfation sites have limited predictive power and objective justification *(9)*. In this study, the composition-based tests traditionally used to predict tyrosine

sulfation have been replaced with the log-odds PSSM, currently the most powerful method for identifying sequence motifs *(7)*. Using a cross-entropy approach *(10)*, the PSSM summarizes the positional effects of amino acids surrounding the tyrosine to determine the probability of a sulfation event. Thus, amino acids common to a specified position near sulfated tyrosine (p-sites) receive high scores while common amino acids in the corresponding position near nonsulfated tyrosine (n-sites) receive low scores.

Our analysis of tyrosine sulfation sites using a statistical evaluation of PSSM-generated sequence scores suggests that TPST demonstrates nonrestrictive substrate specificity. The CDF plot for jackknifed scores and rn-site scores indicates that the overall composition of amino acids, rather than any specific residue, in the sulfation site determines tyrosine sulfation**.**

The observation that certain residues appear with greater frequency within sulfation sites has led to the expectation for a consensus sequence in tyrosine sulfation. Acidic amino acids, particularly at position –1, are frequently found within the sulfation site *(5,6)* and have been the noted consensus feature of sulfation sites. In our set of p-sites, 75% of the sequences have an acidic residue at position –1. Acidic residues may represent a need for a proximal net negative charge and may be particularly effective in exposing the tyrosine residue to the catalytic site of TPST. Cross-entropy analysis verifies that the –1 position has the highest information content, indicating that residues at this position play an important role in determining tyrosine sul–fation. For this reason, it was unusual to find hydrophobic or basic residues at this position in sulfated peptides. The hydrophobic residues phe, ile, and leu in chick antral peptide, coagulation Factor VIII, and platelet glycoprotein 1B, respectively are all found at position –1. Other discrepancies include his found at position –1 in dog fibrinogen and bovine coagulation Factor X precursor, and cys in alpha conotoxin *(11)* and human chorionic gonadotropoin (HCG) *(12)*. A cys residue, in particular, within seven residues of tyrosine *(6)* has long been considered a contraindication for sulfation. In our previous analysis of sulfation sites we noted that cys does not necessarily inhibit tyrosine sulfation, only that cys had not been observed in known sulfated proteins *(9)*. These observations are consistent with the notion that other amino acids, e.g., small charged residues, within the sulfation site can compensate for the presence of unfavorable residues located at the critical –1 position**.**

If amino acids near sulfated tyrosines function mainly to increase the accessibility of the tyrosine instead of serving as specific recognition sites, why are acidic amino acids favored over neutral, polar, or basic amino acids? Neutral and polar amino acids may be less effective in positioning themselves and neighboring residues into a solvent-accessible conformation. The size of basic amino acids, particularly the number of side chain methylene carbons, may hinder accessibility of the tyrosyl side chain. While asp and glu have one and two methylene carbons in their side chains, respectively, the basic amino acids, arg and lys, respectively have three and four nonpolar side chain methylene carbons.

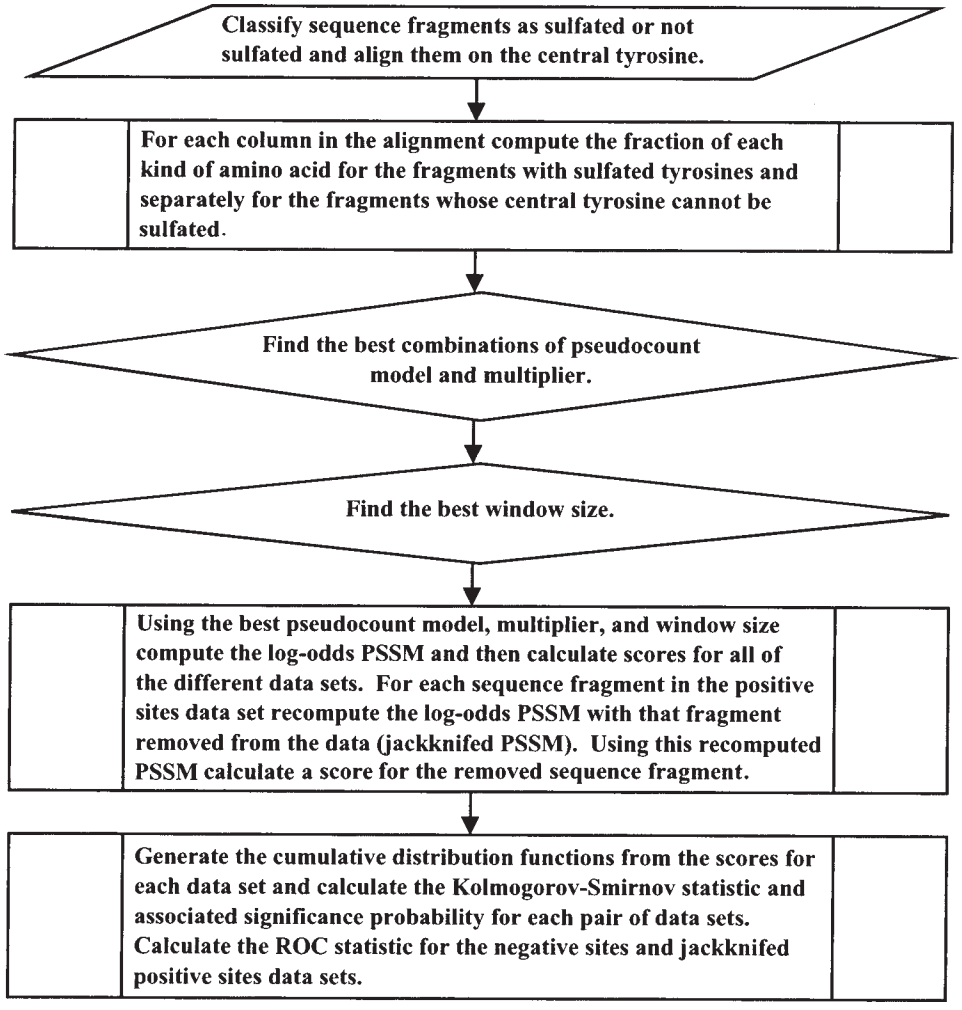
One might expect to find different TPST genes coding for variants of the enzyme that are specific for these differences among sulfation sites. Yet despite extensive searches of the NCBI/GenBank database, only two TPSTs have been found, with the most recently reported TPST of nonmammalian origin *(13)*. Alternatively, one might imagine that TPST instead exhibits a broad range of substrate specificity, dependent only upon the accessibility of the target tyrosine. Thus, residues within the sulfation site serve primarily to position the tyrosine favorably for TPST sulfation. Even the lack of key residues within the sulfation site can be compensated for by other amino acids. CDF plots for jackknifed and rn-site scores in fact show that the entire amino acid composition of the sulfation site dictates the likelihood of sulfation.

Another important function of sulfation may be related to the trafficking of proteins in the cell *(14)*. Twenty of the 44 sulfation sites used to make the PSSM have the sorting signal, YXXO, where Y represents tyrosine, O is any hydrophobic amino acid, and X is any residue. At least two subunits of the clathrin adapator complex involved in targeting membrane proteins to compartments of the endosomal-lysosomal system prefer acidic residues at the –1 positions *(15)*. Because these signals function only when the tyrosine is unmodified, sulfation would nullify the signal for phosphorylation. The fact that tyrosine kinases and sulfotransferases recognize similar sequences *(16)* is further indication of the importance of the sorting sequence in keeping tyrosine phosphorylation and sulfation separate. These phosphorylation sites, associated with signaling pathways, are located in the cell, and do not enter the Golgi secretory pathway. The findings from our study indicate that future analysis of tyrosine sulfation sites may benefit from more extensive comparisons with sorting signals and phosphorylation sites.

The PSSM identified two sequences, horse gastrin and turtle antral peptide, as having a high probability of being sulfated (scoring 4.7 and 17.7 respectively) despite experimental evidence indicating that they are not sulfated in vivo. Interestingly, these peptides are sulfated when expressed in mammalian HIT cells *(8)*. This result suggests that although the primary sequence may be favorable for sulfation, the condition necessary for sulfation may not exist in the cells of the organism in which these proteins are naturally present.

This method of identifying tyrosine sulfation sites has allowed us to identify potential sulfation sites among several candidate hormones, neuropeptides, and cytokines, all of which demonstrate conservation of the site in at least two animals. The criteria for selecting these sites have been based exclusively on log-odds scores from our PSSM, con-

Table 3
Flow Diagram for Creating and Testing the Tyrosine Sulfation PSSM



trasting amino acids present at specific positions around known sulfated tyrosines with those present in the same sites around tyrosines known not to be sulfated. TPST is thus expected to sulfate any tyrosine accessible to its active site provided that the tyrosine is sufficiently exposed and the hydroxyl group is available. This view contrasts with the currently accepted view that there are definite, albeit subtle, recognition features in the amino acid side chains proximal to the sulfated tyrosine.

Materials and Methods

Definition of Known Tyrosine Sulfation Sites and Nonsulfated Sites

All sequences were retrieved from Swiss-Prot release 31. A tyrosine sulfation site was defined as that part of the primary sequence containing the target tyrosine and up to seven flanking amino acids. To avoid overrepresentation of a particular family of sulfation sites in the PSSM, 44 representative sulfation sites were selected from the 103 sites experimentally shown to be sulfated (known sites) with the requirement that members of the same protein family were <70% identical. These sites (tyrosine with a window of ±7) were designated as p-sites. In addition, 293 nonsulfated tyrosines from the same proteins were classified as n-sites. Five hundred random sequences were generated by randomly shuffling the amino acids in the sulfation sites from p-sites, excluding the central tyrosine, and uniformly sampling from this composition to generate an amino acid at each of the 14 nontyrosine sites in the window. These randomly generated sequences were designated rn-sites. A final set of sequences, designated sw-sites, contained all tyrosine sites from the Swiss-Prot database having the keyword metazoa on the Organism Classification (OC) line of the database entry. A flow diagram represents our method of analysis (Table 3).

Constructing the Log-Odds PSSM

To calculate the PSSM, p-site sequences were first aligned around the central tyrosine. The position immedi-

ately N-terminal to the tyrosine is designated –1 and the first C-terminal position is +1. The number and type of amino acid at each position within seven residues of the tyrosine were counted. The performance of the PSSM was improved by increasing the counts for each residue by adding pseudocounts *(17,18)*. In general, pseudocounts improve the PSSM by compensating for the fact that the set of aligned sequences represents a subset of all sulfation sites. Pseudocounts also eliminate zeros in the observed counts, allowing us to apply the log-odds scoring method for sequence motifs identification *(7)*.

Lawrence pseudocounts have the same magnitude and are calculated according to the occurrence of the amino acid within the sulfation site. The Henikoff pseudocounts are based on the frequencies with which amino acids are observed to replace one another in related proteins of the BLOCKS database. These pseudocounts are considered evolutionary in nature and are added in proportion to the number of kinds of amino acids observed at that position of the aligned sequence.

Scores were calculated for each position of the PSSM as follows:

$$\text{Score}\ (i,j) = \log_2 \{\ [c(i,j)/C(i)]\ /\ [n(i,j)/N(i)]\ \}$$

where i is an integer position number in the site from –7 to +7 (N- to C-terminus), j is an integer value from 1 to 20 representing a specific amino acid; $c(i,j)$ is the number of occurrences of amino acid j in column i of the alignment; $C(i)$ is the total number of all amino acids present in column i of the alignment; and $n(i,j)$ and $N(i)$ are corresponding values in the n-sites (the set of background sequences) from which the p-sites are to be distinguished.

In this equation all of the counts, $C(i)$, $c(i,j)$, $N(i)$, and $n(i,j)$ include pseudocounts.

The odds ratio (OR) was calculated by dividing the frequency of each residue at each position in the p-sites by the frequency of each residue at each position in the n-sites. The base 2 logarithm of the OR for each amino acid at a specific position is the score that is part of a PSSM. For any PSSM, the sum of the scores from the PSSM for each amino acid in the sequence being evaluated was used to calculate the score for each potential tyrosine sulfation site. A more positive score suggested that the site was more likely to undergo sulfation.

PSSMs computed from different numbers of aligned sequence positions were compared to determine the optimal window size for discriminating between p-site and n-site sequences. PSSMs based on windows of –2 to +2 through –7 to +7 were evaluated. From ROC analysis, PSSMs made from positions –5 to +5 were the most accurate in classifying sulfation sites in the smallest window size and were used for all subsequent sequence evaluations.

Jackknife Evaluation of the PSSM

The jackknife test was performed on sequences in the p-sites to determine whether the PSSM could identify sulfated protein families not included in its matrix. Jackknifed PSSMs were constructed by using only 43 of the 44 sequences from the p-sites and were used to score the sequences in p-sites. This test was repeated 44 times, once for each sulfation site not included in the PSSM. Since each of the 44 known sulfation sites in the p-sites was unique, its sequence was not explicitly represented by other sites in the jackknife PSSM, and thus their scores mimicked the scores expected for unobserved tyrosine sulfation sites.

Receiver Operating Characteristic

ROC analysis *(19)* was used to test the predictive power of our PSSM. The area under the ROC curve allowed us to evaluate the predictive power of the PSSM to distinguish between two discrete states. In our work, these discrete states were the p-sites, in which the tyrosine is sulfated in a natural biological system, and the n-site, a site where the tyrosine will not be sulfated. PSSM scores for p- and n-sites were sorted in descending order. For a given score, the fraction of p-sites with scores as high or higher were plotted on the y-axis. The fraction of n-sites with scores as high or higher were plotted on the *x*-axis. The maximum ROC score (area under the ROC curve) of one indicated that all p-site sequences scored higher than any n-site sequence. Conversely, a ROC score of zero indicated that all n-site sequences scored higher than any p-site sequence. A ROC value near 0.5 indicated that p-site scores were more or less uniformly intermixed with the n-site sequence scores and that the score provided no basis for discrimination between the cases. Thus, the ROC analysis measured the degree of overlap between a histogram of p-site scores and a histogram of n-sites scores. This was a measure of how likely we were to incorrectly classify a site using a test based on our PSSM.

Evaluation of Different Pseudocount Models

The ROC statistic was used to evaluate the effects of Henikoff or Lawrence pseudocounts with different pseudocount multipliers M, background set, and window sizes (Fig. 4). In general, pseudocounts in an n-site background with a –5 to +5 window yielded the highest ROC statistic and defined the minimum window size that was necessary to determine sulfation based on the primary sequence. The Pattern/Negative and Henikoff/Negative PSSMs in a –5 to +5 window were of virtually identical effectiveness. The pseudocount method and multiplier (Pattern/Negative M = 3 and Henikoff/Negative M = 13) were selected from a point on their respective curves where the increase in the ROC statistic was virtually flat. The ROC statistics for these PSSMs were 0.9763 and 0.9758, respectively. The Pattern/Negative PSSM was chosen based on these results, i.e., minimally better performance.

CDF and the Kolmogorov-Smirnov D Statistic

The CDF plot is a step function that plots the score for the tyrosine site on the horizontal axis and the fraction of scores from that set that have the same or lower score on the

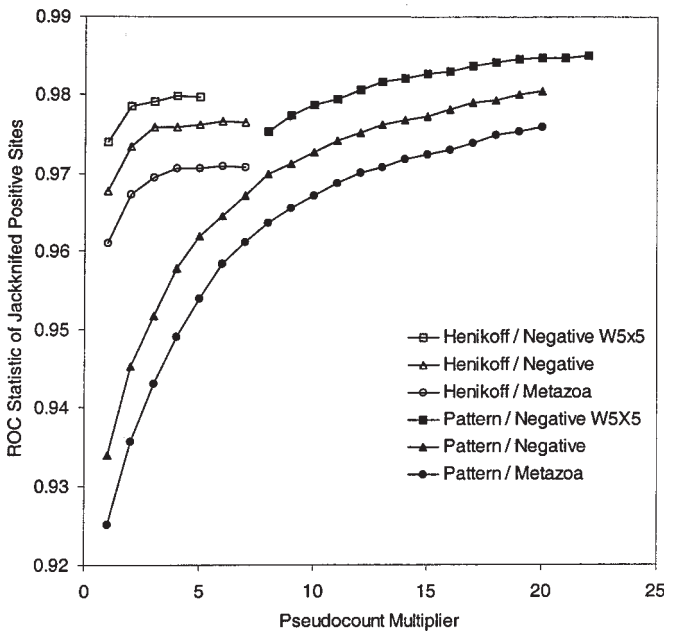


Fig. 4. Selection of the basis of pseudocounts, the pseudocount multiplier M, and window size. The Henikoff/evolutionary pseudocounts and the Lawrence/Pattern p-sites composition pseudocounts were compared in different backgrounds and window sizes. The background set of sequences used were the entire Swiss-Prot Metazoa database (sw-sites) and n-sites. Unless otherwise indicated, the window size was ±7. A higher ROC statistic indicates that a particular combination of pseudocount model and pseudocount multiplier gives more complete discrimination between tyrosines that can be enzymatically sulfated and those that cannot be enzymatically sulfated. A value of 1.0 indicates that all p-sites scored higher than any n-sites.

vertical axis. The Kolmogorov-Smirnov D (KSD) statistic summarizes differences between two sample CDFs and provides a statistical significance level for the differences between them. The KSD statistic is the maximum difference in the fraction of points plotted value for the two CDFs being compared—the maximum vertical separation in the plot. The KSD statistic ranges from 0.0 to 1.0 and is sensitive to multiple aspects of the two CDFs, such as the average, variance, and skewedness. This sensitivity makes it possible to ask more general questions about whether two sets of scores are different than is possible when using only a single aspect of the distribution such as the average. The differences in the scores for the various kinds of sites (p-sites, n-sites, rn-sites, and so on) computed with the same PSSM were summarized with the KSD statistic.

Relative Entropy (Information Content)

The information content *(20)* for each position in the window of –7 to +7 was calculated for the PSSM according to the following equation:

$$I(i) = \Sigma_j \{[c(i,j)/C(i)] \times \log_2 ([c(i,j)/C(i)] / [n(i,j)/N(i)])\}$$

where i is an integer position number from –7 to +7 (N- to C-terminus); j is an integer value from 1 to 20 representing an amino acid; $c(i,j)$ is the number of occurrences of amino acid j in column i of the alignment; $C(i)$ is the total number of all amino acids present in column i of the alignment; and $n(i,j)$ and $N(i)$ are corresponding values in the n-sites (the set of background sequences) from which the p-sites are to be distinguished.

In this equation all of the counts, $C(i)$, $c(i,j)$, $N(i)$, and $n(i,j)$, include pseudocounts.

The information content measures the contribution of each position in the PSSM in discriminating between p- and n-sites and is frequently referred to as the relative or cross entropy of the pair (p- and n-sites) of distributions. Higher information content provides more discriminating power *(10)*. Information content tells us which positions are most important in determining whether the tyrosine will be sulfated.

Acknowledgments

We would like to thank Harriet Lam for help in organizing the data and for validating the tyrosine sulfation sites. This work was supported by funds from grant RR0609 from the National Center for Research Resources.

References

1. Bundgaard, J. R., Vuust, J., and Rehfeld, J. F. (1995). *Embo. Journal* **14,** 3073-3079.
2. Hortin, G. L., Farries, T. C., Graham, J. P., and Atkinson, J. P. (1989) *Proc. Natl. Acad. of Sci. USA* **86,** 1338-1342.
3. Pittman, D. D., Wang, J. H., and Kaufman, R. J. (1992). *Biochemistry* **31,** 3315-3325.
4. Baeuerle, P. A. and Huttner, W. B. (1985). *J. Biol. Chem.* **260,** 6434-6439.
5. Hortin, G., Folz, R. Gordon, J. I., and Strauss, A. W. (1986). *Biochem. Biophys. Res. Commun.* **141,** 326-333.
6. Huttner, W. B. (1988). *Ann. Rev. of Physiol.* **50,** 363-376.
7. Karlin, S. and Altschul, S. F. (1990). *Proc. Natl. Acad. Sci. USA* **87,** 2264-2268.
8. Bundgaard, J. R., Vuust, J., and Rehfeld, J. F. (1997). *J. Biol. Chem.* **272,** 21,700-21,705.
9. Rosenquist, G. L. and Nicholas, H. B., Jr. (1993). *Prot. Sci.* **2,** 215-222.
10. Baldi, P. F. and Brunak, S. (1998). In: Bioinformatics: the machine learning approach. MIT Press: Cambridge.
11. Loughnan, M., Bond, T., et al. (1998). *J. Biol. Chem.* **273,** 15,667-15,674.
12. Bielinska, M. (1987). *Biochem. Biophys. Res. Commun.* **148,**1446-1452.
13. Ouyang, Y. B. and Moore, K. L. (1998). *J. Biol. Chem.* **273,** 24,770-24,774.
14. Trowbridge, I. S. (1993). *Ann. Reviews Cell Bio.* **9,** 129-161.
15. Ohno, H., Aguilar, R. C., Yeh, D., Taura, D., Saito, T., and Bonifacino, J. S. (1998). *J. Biol. Chem.* **273,** 25,915-25,921.
16. Kishimoto, A., Nichiyama, I., and Nakanishi, H. (1985). *J.Biol Chem.* **260,** 12,492-12,499.
17. Henikoff, J. G. and Henikoff, S. (1996). *CABIOS* **12,** 135-143.
18. Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F., and Wootton, J. C. (1993). *Science* **262,** 208-214.
19. Metz, C. E. (1978). *Seminars in Nuclear Medicine* **8,** 283-298.
20. Stormo, G. D. and Hartzell, G. W. (1989). *Proc. Natl. Acad. Sci. USA* **86,**1183-1187.