Vol. 18 no. 5 2002
Pages 769–770

The Sulfinator: predicting tyrosine sulfation sites in protein sequences

Flavio Monigatti, Elisabeth Gasteiger*, Amos Bairoch and Eva Jung

SWISS-PROT Group, Swiss Institute of Bioinformatics, CMU, 1 rue Michel-Servet, CH-1211 Geneva 4, Switzerland

Received on August 29, 2001; revised on December 3, 2001; accepted on December 7, 2001

ABSTRACT

Summary: Protein tyrosine sulfation is an important post-translational modification of proteins that go through the secretory pathway. No clear-cut acceptor motif can be defined that allows the prediction of tyrosine sulfation sites in polypeptide chains. The Sulfinator is a software tool that can be used to predict tyrosine sulfation sites in protein sequences with an overall accuracy of 98%. Four different Hidden Markov Models were constructed, each of them specialized to recognize sulfated tyrosine residues depending on their location within the sequence: near the N-terminus, near the C-terminus, in the center of a window with a size of at least 25 amino acids, as well as in windows containing several tyrosine residues.

Availability: The Sulfinator is accessible at (http://www.expasy.org/tools/sulfinator/).

Supplementary information: Sulfinator documentation is accessible at (http://www.expasy.org/tools/sulfinator/sulfinator-doc.html).

Abbreviations: SWP: SWISS-PROT accession number.

INTRODUCTION

Proteins, once synthesized on the ribosomes, are subject to a multitude of modification steps in order to be fully functional (Han and Martinage, 1992; Jung *et al.*, 2001). They are cleaved (e.g. eliminating signal sequences or initiator methionines); many simple chemical groups can be attached to them (e.g. acetyl, methyl, phosphoryl, etc.) as well as some more complex molecules, such as sugars and lipids. Finally, they can be internally or externally cross-linked (e.g. disulfide bonds). Taking into account alternative splicing of mRNA as well, the number of different protein molecules expressed by for instance the human genome is probably closer to a million than to the 30 000 to 60 000 generally considered by genome scientists.

Tyrosine sulfation is a ubiquitous post-translational modification of proteins that go through the secretory pathway within metazoic cells. While the exact biological role of tyrosine sulfation is largely unknown, it has been shown that tyrosine sulfation increases the hemolytic activity of complement C4 (SWP: P01028), it plays an important role in receptor binding of the peptide hormone cholecystokinin (SWP: P06307) and it increases interactions between von Willebrand factor and Factor VIII (SWP: P00451) as well as between hirudin and thrombin (SWP: P01050).

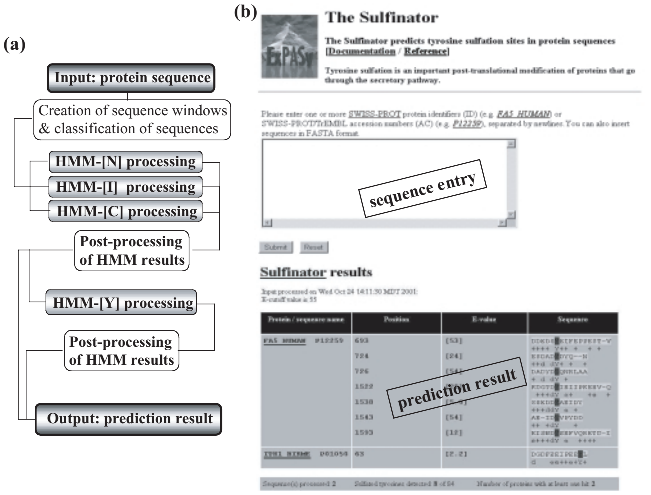
Although the tyrosylprotein–sulfotransferases (TPSTs, EC 2.8.2.20) have long been identified in various metazoa, no clear-cut acceptor sequence could be defined that allows prediction of tyrosine sulfation events in these organisms. So far, only a rule for protein tyrosine sulfation has been described in PROSITE (http://www.expasy.org/cgi-bin/nicesite.pl?PS00003). This rule describes several favoured and less favoured amino acids around potential target tyrosine residues, however, does not allow the automated screening of protein sequences for tyrosine sulfation sites as the rules are not specific enough. This was tested internally by implementing the rules in a perl script.

METHODS AND IMPLEMENTATION

The Sulfinator (Figure 1a) is a software tool that can be used to predict tyrosine sulfation sites in protein sequences. It employs four different Hidden Markov Models (HMMER software package; Eddy, 1998) that were built to recognize sulfated tyrosine residues located N-terminally (HMM-[N]), within sequence windows of more than 25 amino acids (HMM-[I]) and C-terminally (HMM-[C]) as well as sulfated tyrosine residues clustered within 25 amino acid windows (HMM-[Y]), respectively. In the following we will briefly explain the parameters used to construct the HMMs, however, for more in depth explanations of these parameters, we would like to refer to the excellent HMM user's guide (http://hmmer.wustl.edu/).

All four HMMs contain information from one multiple sequence alignment and implement the

*To whom all correspondence should be addressed.



The Sulfinator

The Sulfinator predicts tyrosine sulfation sites in protein sequences [Documentation / Reference]

Tyrosine sulfation is an important post-translational modification of proteins that go through the secretory pathway.

Please enter one or more SWISS-PROT protein identifiers (ID) (e.g. FA5_HUMAN) or SWISS-PROT/TrEMBL accession numbers (AC) (e.g. P12259), separated by newlines. You can also insert sequences in FASTA format.

Sulfinator results

Input processed on Wed Oct 24 14:11:30 MDT 2001:
E-cutoff value is 55

Protein / sequence name	Position	E-value	Sequence
FA5_HUMAN P12259	693	[53]	[illegible]
	724	[24]	[illegible]
	726	[54]	[illegible]
	1522	[illegible]	[illegible]
	1530	[5.4]	[illegible]
	1543	[54]	[illegible]
	1593	[12]	[illegible]
ITH1_HIRME P01050	63	[2.2]	[illegible]

Sequence(s) processed: 2 Sulfated tyrosines detected: 8 of 54 Number of proteins with at least one hit: 2

Fig. 1. The general architecture of the Sulfinator is shown in a flowchart (a). Screenshots from the Sulfinator web page illustrate the query form with input options available for the user (top) and the format of an output is shown (bottom) (b).

maximum weight algorithm (Krogh and Mitchison, 1995) for 'weighting' sequences in the multiple sequence alignment as well as the *a priori* mixture Dirichlet method. Detailed information on data sets used for this project can be found online (http://www.expasy.org/tools/sulfinator/sulfinator-doc.html).
The main HMM chain, HMM-[I], was built to recognize tyrosine sulfation sites within sequence windows centred around target tyrosines and was configured to find a single global alignment to the target sequence. Although sequence windows of 23 amino acids were used to construct HMM-[I], the best performing linear HMM chain only consisted of 17 amino acids, 5 N-terminally and 11 C-terminally to the target tyrosine. This led to the assumption that only 5 amino acids N-terminally to a target tyrosine seem to contain useful information to predict tyrosine sulfation events. HMM-[N], HMM-[C] and HMM-[Y] (Figure 1a) were, in contrast to HMM-[I], configured to find multiple domains per sequence, where each domain can be a local alignment. While constructing these HMMs, the '–gapmax' value was lowered to 0.2, in order to control the '–fast' model construction algorithm, which was used for all four HMMs. The reduction of the '–gapmax' value ensures that fewer columns get assigned to the consensus and the models get smaller. To gain more independent control over the local alignment behaviour of HMM-[N] and HMM-[C], the '–swentry' and '–swexit' options were included, respectively. In case of HMM-[N] the probability value of '–swentry' was decreased to 0.2, meaning that matching to HMM-[N] will not be forced to begin with the start state of HMM-[N]. However, once an alignment has started, it must continue over the full HMM length. The opposite situation can be found for HMM-[C], where the '-swexit' value was slightly increased to 0.7 to allow termination of an alignment to HMM-[C] at the C-terminus of a protein sequence. But, alignments have to start with the start state of HMM-[C]. In case of HMM-[Y], the '–swentry' and '–swexit' values were set on 0, to strongly penalize matching of fragments. In addition, the 'architecture prior'('–archpri') value was slightly increased to 0.9 in order to govern geometric prior distribution over HMM length and slightly favour longer models. All four HMMs are combined in the Sulfinator (Figure 1b), a perl script. The web interface to the Sulfinator is accessible on the ExPASy server at the URL http://www.expasy.org/tools/sulfinator/.

CONCLUSIONS

Using independent test sets containing sequences with validated non-sulfated and sulfated tyrosine residues, the Sulfinator correctly predicts 98% of the tyrosine sulfation sites and 98% of the non-sulfated tyrosine residues (test sets are included in on-line documentation). To get an idea on how many proteins in a proteome might be sulfated, we scanned proteins from *Homo sapiens*, *Drosophila melanogaster* and *Caenorhabditis elegans* that are described in SWISS-PROT to go through the secretory pathway. The results from these proteome scans suggested that in each of these organisms one third of the proteins that enter the secretory pathway may contain on average two tyrosine sulfation sites per protein.

We believe that in the era of proteomics, more emphasis will be placed on the characterization of post-translational modifications, and that bioinformatics tools for the prediction of specific post-translational modifications will be very useful in assisting the discovery of these important protein modifications.

REFERENCES

Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

Han,K.K. and Martinage,A. (1992) Post-translational chemical modification(s) of proteins. *Int. J. Biochem.*, **24**, 19–28.

Jung,E., Veuthey,A.-L., Gasteiger,E. and Bairoch,A. (2001) Annotation of glycoproteins in the SWISS-PROT database. *Proteomics*, **1**, 262–268.

Krogh,A. and Mitchison,G. (1995) Maximum entropy weighting of aligned sequences of proteins or DNA. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **3**, 215–221.