Biases and Complex Patterns in the Residues Flanking Protein N-glycosylation Sites

Rubin, E.[1], Ben-Dor, S.[1] and Sharon, N.[2]

[1] Bioinformatics and Biological Computing Unit, Biological Services, Weizmann Institute of Science

[2] Department of Biological Chemistry, Weizmann Institute of Science

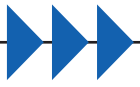
Protein glycosylation, in particular of asparagine residues (N-glycosylation) is the most common and most complex reaction that occurs during protein biosynthesis and often affects markedly their physicochemical and biological properties. It has been estimated that over half of proteins in Nature are glycoproteins (Apweiler et al., 1999). The consensus for N-glycosylation, also known as the sequon, is NXT/S; it is abundant in proteins, but only two thirds are glycosylated. The lack glycosylation of some sequons may be a result, at least in part, of the presence or absence of specific residues at or near the sequon. For example, a proline (Pro) at the X position was reported to be prohibitive for glycosylation. Little is known, however, about the influence of other residues at this position, nor of those flanking the sequon, on the efficiency of N-glycosylation (Shakin-Eshelman, 1996).

We extended traditional approaches of sequence analysis to glycosylation sites in several ways. Using the current version of SWISSPROT, in which 602 well characterized, non-redundant N-glycoproteins have been deposited. The analyzed pattern was extended from the traditional 3-mer sequon NXS/T to a 7-mer sequon M_2M_1NXS/TP_1P_2. Based on experimental information on N-glycosylation of specific aspargines deposited in SWISSPROT, 1186 glycosylated and 717 non-glycosylated 7-mer sequons were analyzed. A supervised learning approach was used to identify complex patterns that separate glycosylated and non-glycosylated sequons.

Analysis of the amino acid distribution at each position of the 7-mer sequon revealed biases in all. Glycosylated sequons showed over-representation of Gly in the X position, and of Leu in the P_1 position and under-representation of Pro in these positions. For non-glycosylated sequons, over-representation of Ser was found in M_2, Asp in M_1, Lys and Pro in X, Tyr in P_1 and Gly in P_2; under-representation of Leu was observed in position P_1.

Supervised learning identified two complex patterns. The data-mining tool WizWhy (WizSoft, Israel) was used to analyze the 7-mer sequons, by describing each position as a separate attribute, and providing the glycosylation state of each sequon as the dependent variable. WizWhy identifies complex "rules" or patterns by first identifying biases in single sites, and merging "rules" that together better explain the dependent attribute.

In glycosylated sequons, several sub-patterns were identified, all matching the consensus D/ESNGTLT. Each sub-pattern matched 2-3 amino acids in positions $M_{1\text{-}2}$, X, and $P_{1\text{-}2}$ of the consensus. Scanning SWISSPROT for sequences



over-represented oligomers, only those exhibiting ratios much higher than the cutoff were selected as regulatory elements. We identified the GCN4 binding site with an extremely high frequency ratio of 10.00 whereas the cutoff was 5.00. GCN4 is a master gene involved in amino-acid metabolism and has already been reported as central in the over-expression of amino-acid metabolism genes during cell starvation (Marton MJ, 2001). All yeast genes that are well known to be regulated by GCN4 (e.g. HIS4, TRP4, ILV1, ILV2, ADE4, ARG1, ARG8, HIS3). were also over-expressed as expected. Another regulatory element we isolated was the CPF1 binding site. CPF1 is involved in the regulation of methionine synthesis. All genes that are regulated by this element (e.g. PGK1, CYT1, and MET16) were also over expressed as expected. We performed the same analysis on a group of genes that participate in the carbohydrate metabolism , and identified oligomers that exhibited ratios far above the cutoff (e.g. GGGGACT ,GGGGG, GGGGC and CGCGC). No literature was found regarding those oligomers. We therefore propose them as regulatory elements of the carbohydrate metabolism during cell starvation.

Discussion Our method has succeeded in identifying some important regulatory elements involved in the process of gene over-expression during cell starvation by analysis of data derived from a microarray .This enables us to throw light on the regulatory mechanism underlying the over-expression phenomenon by suggesting some of the elements involved. Our encouraging results are due to two aspects: First, we focus on genes that react quickly to a certain change (e.g. elevation of rapamycin in the cell media) because this suggests they are responding to a relatively specific common cause. This common cause is likely to co-occur with a small over-represented group of regulatory elements suggesting that our method is effective in these cases. Second, we use statistical simulation to derive a cutoff value, reducing the number of false-positives to below 5\%, making our method very reliable.

References

[1] Hardwick J.S., Kuruvilla F.G., Tong J.K., Shamji A.F. and Schreiber S.L.
Rapamycin-modulated transcription defines the subset of nutrient-sensitive signaling pathways directly controlled by the Tor proteins. Proc. Natl. Acad. Sci USA, 1999 Dec 21;96.

[2] Hughes J.D., Estep P.W., Tavazoie S., Church G.M .Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. J.Mol.Biol.2000, Mar 10;296(5):1205-14.

[3] Natarajan K., Meyer M.R., Jackson B.M., Slade D., Roberts C. Transcriptional profiling shows that Gcn4p is a master regulator of gene expression during amino acid starvation in yeas. Hinnebusch A.G., Marton M.J. J. Mol. Cell Biolology 2001, 21(13):4347-68.

[4] van Helden J., Andre B., Collado-Vides J. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. J.Mol. Biology 1998, 281(5):827-42.

e-mail: chaya@cs.technion.ac.il,
e-mail: dang@cs.technion.ac.il