Sequence differences between glycosylated and non-glycosylated Asn-X-Thr/Ser acceptor sites: implications for protein engineering

Ylva Gavel and Gunnar von Heijne

Research Group for Theoretical Biophysics, Department of Theoretical Physics, Royal Institute of Technology, S-100 44 Stockholm and

Department of Molecular Biology, Karolinska Institute, Center for Biotechnology, NOVUM, S-141 52 Huddinge, Sweden

In N-glycosylated glycoproteins, carbohydrate is attached to Asn in the sequence Asn-X-Ser/Thr, where X denotes any amino acid. However, the presence of this consensus peptide does not always lead to glycosylation. We have compiled an extensive collection of glycosylated and non-glycosylated Asn-X-Thr/Ser sites and present a statistical study based on this data set. Our results indicate that non-glycosylated sites tend to be found more frequently towards the C terminus of glycoproteins, and that proline residues in positions X and Y in the consensus Asn-X-Thr/Ser-Y strongly reduce the likelihood of N-linked glycosylation. Beyond this, there are no obvious local sequence features that seem to correlate with the absence or presence of N-linked glycosylation. These findings are discussed in terms of the prediction and engineering of glycosylation sites in secretory proteins.

Key words: glycoproteins/N-glycosylation/statistical study

Introduction

The attachment of N-linked carbohydrates to proteins is thought to occur during or shortly after translocation of the nascent chain into the lumen of the endoplasmic reticulum (Kaplan et al., 1987; Lennarz, 1988; Hubbard and Ivatt, 1981). The oligosaccharide chain is transferred by the enzyme oligosaccharyl transferase to the asparagine in the consensus tripeptide Asn-X-Thr/Ser, where X is any amino acid (Marshall, 1972). Most putative acceptor sites that become exposed on the luminal side of the endoplasmic reticulum (ER) membrane are efficiently glycosylated, but some are never used. It has long been known that glycosylation is blocked when X is a proline (Mononen and Karjalainen, 1984), but this rule accounts for only a minor portion of all known non-glycosylated consensus sites.

Here, we present a study based on a data set of carefully selected glycosylated (gs+) and non-glycosylated (gs–) Asn-X-Thr/Ser sites. All gs– sites included in this set have been checked in the literature. Sites from homologous proteins have been removed, both from the gs+ and gs– sets. Statistical methods have been used in order to test the significance of the results.

Our analysis indicates that glycosylation is strongly inhibited by proline residues both in positions X and Y in the consensus Asn-X-Thr/Ser-Y. Also, non-glycosylated sites tend to be found more frequently towards the C terminus of the proteins in our sample, whereas glycosylated sites are rare in this region. These observations allow prediction of gs+ sites to be made with ~95% confidence, whereas only some 25% of all gs– sites can be reliably predicted from the primary sequence.

Materials and methods

Sequences of glycoproteins were collected from the literature and from the NBRF-PIR database (George et al., 1986).

Preparation of data sets

Earlier studies of the sequence patterns associated with N-linked glycosylation suffer from a number of methodological shortcomings, e.g. small sample sizes, inclusion of sites from homologous proteins, no statistical analysis. But the most obvious weakness is that too little attention has been paid to the collection of proper sets of both gs+ and gs– sites, thus precluding any useful comparisons between the two classes of sequences.

gs+ sites must be picked with some care. Many proteins are non-glycosylated merely because they are never exposed to the carbohydrate-transferring enzyme. Asn-X-Thr/Ser sequences from such proteins should not be included in the gs+ set. Since N-glycosylation is thought to occur in the lumen of the ER (Kaplan et al., 1987; Lennarz, 1988), gs– sites from cytoplasmic proteins must be excluded. The same is true of sites from intracellular and transmembrane parts of membrane proteins. Furthermore, sites from proteins produced by cells unable to carry out N-glycosylation must be avoided. For these reasons, we have restricted the data set to gs– sites found in luminal domains within the sequences of proteins that also contain gs+ sites and thus are certain to have been exposed to the oligosaccharyl transferase.

In an extensive literature search, we found a total of 55 gs+ sites in proteins known to be N-glycosylated. A total of 48 gs+ sites were included in our final data set. The rest (Robinson and Appella, 1979; Takahashi et al., 1984; Van Den Berg et al., 1976, 1977; Beintema, 1985; Havinga and Beintema, 1980) were from highly homologous proteins, mainly in pancreatic ribonucleases.

We also collected ~600 gs– sites from the NBRF database and from the literature. Again, obviously homologous proteins were removed in order to avoid distortions of the statistics. The final version of our data set contained 417 gs– sites.

All the gs+ sites were explicitly stated to be glycosylated in the literature or in the database. Some references reported non-glycosylated sites as well. For the rest of the gs– sites, the absence of sugar could be inferred from experimental data presented in the literature. Some potential glycosylation sites were located in carbohydrate-free tryptic peptides and therefore were not glycosylated. In other cases, it was possible to make assignments based on the results from sequence determinations.

With most sequencing methods, a glycosylated residue cannot be detected; instead, a blank appears in the sequence, and the amino acid in this position has to be identified by other means. Therefore, if some of the asparagines found in the Asn-X-Thr/Ser sequences of a protein show up as blanks whereas others do not, those which give an Asn signal can be assumed to be non-glycosylated.

The sites included in the data set are given in Table I. In some
| Proteins containing both gs⁺ and gs⁻ sites                                      | gs⁺ | gs⁻ | Experimental evidence | Reference                      |
|--------------------------------------------------------------------------------|-----|-----|------------------------|--------------------------------|
| Alkaline extracellular protease (Y.lipolytica)                                 | 123 | 292 | 1                      | Matoba et al. (1988)           |
|                                                                                | 330 | 5   |                        |                                |
| Alkaline phosphatase (human placenta)                                         | 249 | 122 | 2                      | Millan (1986)                  |
| Alpha-antitrypsin (human)                                                     | 46  | 83  | 1                      | Carrell et al. (1982)          |
|                                                                                | 247 | 1   |                        | Mega et al. (1980)             |
|                                                                                | 390 | 5   |                        | Carrell et al. (1981)          |
| Alpha-lactalbumin (rabbit)                                                    | 45  | 84  | 1                      | Hopp and Woods (1979)          |
| Alpha-lactalbumin (bovine)                                                    | 42* | 73  | 3                      | Hopp and Woods (1979)          |
| Aspartic proteinase (Rhizomucor miehei)                                       | 79  | 1   |                        | Struck et al. (1978)           |
|                                                                                | 188 | 1   |                        | Boel et al. (1986)             |
|                                                                                | 313 | 3, 5|                        |                                |
| Beta-hexosaminidase beta-chain (human)                                        | 142 | 1   | 4                      | Mahuran et al. (1988)          |
|                                                                                | 323 | 4   |                        |                                |
| Cathepsin D (porcine spleen)                                                  | 70  | 1   |                        | Shewale and Tang (1984)        |
|                                                                                | 192 | 1   |                        | Takahashi and Tang (1983)      |
| Ceruloplasmin (human)                                                         | 119 | 2   | 2                      | Takahashi et al. (1984)        |
|                                                                                | 339 | 2   |                        |                                |
|                                                                                | 378 | 2   |                        |                                |
|                                                                                | 569 | 3   |                        |                                |
|                                                                                | 743 | 2   |                        |                                |
|                                                                                | 907*| 3   |                        |                                |
| Cholinesterase (human serum)                                                  | 17  | 1   |                        | Lockridge et al. (1987)        |
|                                                                                | 57  | 1   |                        |                                |
|                                                                                | 106 | 1   |                        |                                |
|                                                                                | 241 | 1   |                        |                                |
|                                                                                | 256 | 1   |                        |                                |
|                                                                                | 341 | 1   |                        |                                |
|                                                                                | 455 | 1   |                        |                                |
|                                                                                | 481 | 1   |                        |                                |
|                                                                                | 485 | 3   |                        |                                |
|                                                                                | 486 | 1   |                        |                                |
| Deoxyribonuclease (bovine pancreas)                                           | 18  | 1   |                        | Liao et al. (1973)             |
|                                                                                | 103 | 5   |                        | Catley et al. (1969)           |
|                                                                                | 313 | 2   |                        | Salnikow et al. (1973)         |
| Epidermal growth factor receptor (human)                                     | 104 | 2   |                        | Ullrich et al. (1984)          |
|                                                                                | 151 | 2   |                        |                                |
|                                                                                | 172 | 2   |                        |                                |
|                                                                                | 247 | 4   |                        |                                |
|                                                                                | 328 | 1, 2|                        |                                |
|                                                                                | 337 | 2   |                        |                                |
|                                                                                | 389 | 2   |                        |                                |
|                                                                                | 420 | 2   |                        |                                |
|                                                                                | 504 | 2   |                        |                                |
|                                                                                | 543 | 2   |                        |                                |
|                                                                                | 579 | 2   |                        |                                |
|                                                                                | 590 | 2   |                        |                                |
| Folate-binding protein (cow's milk)                                          | 49  | 1   |                        | Hansen et al. (1984)           |
|                                                                                | 141 | 1   |                        |                                |
|                                                                                | 181 | 4   |                        |                                |
|                                                                                | 212 | 4, 5|                        |                                |
| Folllistatin (porcine)                                                        | 259 | 1   |                        | Shimasaki et al. (1988)        |
| Glucoamylase G1 (Aspergillus niger)                                           | 169 | 1   |                        | Esch et al. (1987)             |
|                                                                                | 180 | 3, 4, 6|                   | Svensson et al. (1983)         |
|                                                                                | 393 | 1   |                        |                                |
Table I. Continued

| Proteins containing both gs+ and gs− sites                                | gs+ | gs− | Experimental evidence | Reference                     |
|-------------------------------------------------------------------------|-----|-----|------------------------|--------------------------------|
| Glycophorin (porcine erythrocyte)                                       | 19  | 16  | 4                      | Honma et al. (1980)            |
|                                                                         | 39  | 1   |                        |                                |
| Immunoglobulin A, alpha (heavy) chain:                                   |     |     |                        |                                |
| (mouse myeloma J558, MOPC 511)                                          | 162 | 129 | 5                      | Taylor and Wall (1988)         |
|                                                                         | 223 | 1   | 3, 5                   | Tucker et al. (1988)           |
|                                                                         | 402 | 3   | 4                      | Robinson and Appella (1980)    |
|                                                                         | 419 | 1   |                        | Frangone et al. (1980)         |
| Immunoglobulin G, gamma3 (heavy) chain (human deletion mutant Wis)      | 6   | 1   | 2                      | Kehry et al. (1979)            |
|                                                                         | 140 | 2   |                        | Kehry et al. (1982)            |
|                                                                         | 235 | 5   |                        |                                |
| Immunoglobulin M, mu (heavy) chain (murine myeloma MOPC 104E)           | 57  | 1   |                        | Shimizu et al. (1971)          |
|                                                                         | 171 | 1   |                        | Putnam et al. (1973)           |
|                                                                         | 332 | 1   |                        | Aubert et al. (1976)           |
|                                                                         | 364 | 1   |                        |                                |
|                                                                         | 402 | 1   |                        |                                |
|                                                                         | 563 | 1   |                        |                                |
| Immunoglobulin M, mu (heavy) chain (human Ou, Waldenstroms)             | 170 | 42  | 3                      | Gebhard et al. (1988)          |
|                                                                         | 332 | 2   |                        |                                |
|                                                                         | 395 | 1   |                        |                                |
|                                                                         | 402 | 1   |                        |                                |
|                                                                         | 563 | 1   |                        |                                |
| Inter-alpha-trypsin inhibitor, 2nd component (human)                     | 64  | 1   | 2                      | Metz-Boutige et al. (1981)     |
|                                                                         | 617 | 2   |                        |                                |
| Lactotransferrin                                                         | 81  | 4   | 5                      |                                |
|                                                                         | 167 | 1   |                        |                                |
|                                                                         | 268 | 4   |                        |                                |
| Leucine-rich alpha2-glycoprotein (human)                                | 44  | 2   |                        | Takahashi et al. (1985)        |
|                                                                         | 151 | 2   |                        |                                |
|                                                                         | 234 | 2   |                        |                                |
|                                                                         | 271 | 3   |                        |                                |
| Lipase (porcine pancreas)                                               | 166 | 1   | 2                      | Blanchetta et al. (1979)       |
|                                                                         | 409 | 4   |                        | De Caro et al. (1981)          |
| NCA-50 (human)                                                          | 118 | 1   |                        | Grunert et al. (1988)          |
|                                                                         | 139 | 1   |                        |                                |
|                                                                         | 254 | 3   | 4                      |                                |
| Ovalbumin (hen)                                                         | 292 | 1   | 2                      | Nisbet et al. (1981)           |
|                                                                         | 311 | 3   |                        |                                |
| Ovotransferrin (hen)                                                    | 473 | 1   |                        | Williams et al. (1982)         |
|                                                                         | 618 | 5   |                        | Kingston and Williams (1975)   |
|                                                                         | 672 | 5   |                        |                                |
| Pepsinogen (hen)                                                        | 113 | 1   |                        | Baudys and Kostka (1983)       |
|                                                                         | 218 | 3   |                        | Welinder (1976)                |
| Peroxidase (horseradish)                                                | 13  | 1   |                        |                                |
|                                                                         | 57  | 1   |                        |                                |
|                                                                         | 158 | 1   |                        |                                |
|                                                                         | 186 | 1   |                        |                                |
|                                                                         | 198 | 1   |                        |                                |
|                                                                         | 214 | 1   |                        |                                |
|                                                                         | 255 | 1   |                        |                                |
|                                                                         | 268 | 1   |                        |                                |
|                                                                         | 286 | 3   |                        |                                |
| Ribonuclease (coypu pancreas)                                           | 34  | 1   | 3                      | Van Den Berg et al. (1976)     |
|                                                                         | 94  | 3   |                        |                                |
| Ribonuclease B (guinea-pig pancreas)                                    | 21  | 1   | 3                      | Van Den Berg et al. (1977)     |
|                                                                         | 34  | 1   |                        |                                |
|                                                                         | 94  | 3   |                        |                                |
### Table 1. Continued

| Proteins containing both gs⁺ and gs⁻ sites | gs⁺ | gs⁻ | Experimental evidence | Reference |
|-------------------------------------------|-----|-----|-----------------------|-----------|
| Ribonuclease (hippopotamus pancreas)      | 34⁺ | 21  | 3                     | Havinga and Beintema (1980) |
|                                           | 34ᵇ | 3   | 1                     |           |
|                                           | 62  | 3   |                       |           |
|                                           | 76  | 3   |                       |           |
| Ribonuclease (horse pancreas)             | 22ᵃᵇ| 22⁺ | 3                     | Beintema (1985) |
|                                           | 34⁺ | 1   |                       |           |
|                                           | 62  | 1   |                       |           |
| S8-glycoprotein (Brassica campestris)⁵    | 15  | 1   |                       | Takayama et al. (1987) |
|                                           | 33  | 1   |                       |           |
|                                           | 83  | 1   |                       |           |
|                                           | 90  | 1   |                       |           |
|                                           | 119 | 1   |                       |           |
|                                           | 214 | 1   |                       |           |
|                                           | 230 | 1   |                       |           |
|                                           | 284 | 3, 4|                       |           |
|                                           | 359 | 3, 4|                       |           |
| Sex steroid binding protein (human)       | 351 | 1   |                       | Walsh et al. (1986) |
|                                           | 367 | 1   |                       |           |
| Thy-1 glycoprotein (rat brain)            | 23  | 1   |                       | Campbell et al. (1981) |
|                                           | 74  | 1   |                       |           |
|                                           | 93  | 3   |                       |           |
| Tissue-type plasminogen activator (human) | 117 | 2   |                       | Hansen et al. (1988) |
|                                           | 184 | 2   |                       | Pennica et al. (1983) |
|                                           | 218 | 3   |                       |           |
| von Willebrand factor (human)             | 94  | 1   |                       | Titani et al. (1986) |
|                                           | 364 | 1   |                       |           |
|                                           | 452 | 3, 4|                       |           |
|                                           | 468 | 1   |                       |           |
|                                           | 752 | 1   |                       |           |
|                                           | 811 | 1   |                       |           |
|                                           | 1460| 1   |                       |           |
|                                           | 1527| 1   |                       |           |
|                                           | 1594| 1   |                       |           |
|                                           | 1637| 1   |                       |           |
|                                           | 1783| 1   |                       |           |
|                                           | 1822| 1   |                       |           |
|                                           | 1872| 3, 4|                       |           |
|                                           | 2027| 1   |                       |           |

| Proteins containing gs⁺ sites only        | gs⁺ sites excluded due to homology | Reference          |
|-------------------------------------------|------------------------------------|--------------------|
| 120 kd lysosomal membrane glycoprotein (rat)|                                    | Howe et al. (1988) |
| 7s nerve growth factor gamma chain (mouse)|                                    | Lewis et al. (1985) |
| Alpha1-acid glycoprotein (human)           |                                    | NBRF               |
| Alpha1-B glycoprotein (human)              |                                    | NBRF               |
| Alpha2-macroglobulin (human)               |                                    | NBRF               |
| Antithrombin-III (human)                   |                                    | NBRF               |
| Apolipoprotein-III (migratory locust)      |                                    | Kanost et al. (1988) |
| Apolipoprotein D (human)                   |                                    | Drayna et al. (1986) |
| Avidin (chicken)                           |                                    | NBRF               |
| Beta2-glycoprotein I (human)               |                                    | NBRF               |
| Biliary glycoprotein I (human)             |                                    | Hinoda et al. (1988) |
| C1 inhibitor (human)                       |                                    | Bock et al. (1986)  |
| Calcequinarin (rabbit)                     |                                    | Fliegel et al. (1987) |
| Carboxypeptidase Y (yeast)                 |                                    | Svensen et al. (1982) |
| Cathepsin B (rat)                          |                                    | NBRF               |
| Cathepsin H (rat)                          |                                    | NBRF               |

436
### Table I. Continued

| Proteins containing $\alpha^+\beta^+$ sites only | $\alpha^+\beta^+$ sites excluded due to homology | Reference |
|-----------------------------------------------|-----------------------------------------------|-----------|
| Cathepsin L (rat)                            |                                               | Ishidoh et al. (1987) |
| Chorionic gonadotrophin $\alpha$-subunit (human) |                                               | Morgan et al. (1975) |
| Chorionic gonadotrophin $\beta$-subunit (human) |                                               | Morgan et al. (1975) |
| Coagulation factor VII (bovine)               |                                               | Takeda et al. (1988) |
| Coagulation factor IX (bovine)                |                                               | NBRF       |
| Coagulation factor X (bovine)                 |                                               | NBRF       |
| Coagulation factor XI (human)                 |                                               | NBRF       |
| Coagulation factor XIIa (human)               |                                               | McMullen and Fujikawa (1985) |
| Collagen II alpha1 chain (chicken)            |                                               | Fujikawa and McMullen (1983) |
| Collagen IV alpha1-subunit (human)            |                                               | Ishidoh et al. (1987) |
| Collagenase (human)                          |                                               | Morgan et al. (1975) |
| Complement component Clq A-chain (human)      |                                               | Morgan et al. (1975) |
| Complement component C1r A-chain (human)      |                                               | Morgan et al. (1975) |
| Complement component C3 (human)               |                                               | Morgan et al. (1975) |
| Complement component C5                       |                                               | Morgan et al. (1975) |
| Complement factor B (human)                   |                                               | Morgan et al. (1975) |
| C-reactive protein 1.1 (Limulus polyphemus)   |                                               | Morgan et al. (1975) |
| Erythropoietin (human)                        |                                               | Morgan et al. (1975) |
| Fibrin beta-subunit (Vicia faba)              |                                               | Morgan et al. (1975) |
| Fibronogen beta chain (human)                 |                                               | Morgan et al. (1975) |
| Fibronogen gamma-A chain (human)              |                                               | Morgan et al. (1975) |
| Fibronogen gamma chain (lamprey)              |                                               | Morgan et al. (1975) |
| Fibronopeptide B (sea lamprey)                |                                               | Morgan et al. (1975) |
| Glucosylceramide activator protein (human)    |                                               | Morgan et al. (1975) |
| Glycophorin A (human)                         |                                               | Morgan et al. (1975) |
| Glycophorin C (human)                         |                                               | Morgan et al. (1975) |
| GluT1A protein (Friend leukaemia virus)       |                                               | Morgan et al. (1975) |
| H-2Db class I antigen alpha chain (mouse)     |                                               | Morgan et al. (1975) |
| Haemagglutinin HA1 chain (influenza)          |                                               | Morgan et al. (1975) |
| Haemagglutinin HA2 chain (influenza)          |                                               | Morgan et al. (1975) |
| Haptoglobin-1 (human)                         |                                               | Morgan et al. (1975) |
| Haptenic a chain (lobster)                    |                                               | Morgan et al. (1975) |
| Haptoglobin (human)                           |                                               | Morgan et al. (1975) |
| Hepatic lectin (chicken)                      |                                               | Morgan et al. (1975) |
| HLA class I histocompatibility antigen alpha chain (human) | | Morgan et al. (1975) |
| HLA class II histocompatibility antigen DC $\beta$-chain (human) | | Morgan et al. (1975) |
| HLA class II histocompatibility antigen DR $\alpha$-chain (human) | | Morgan et al. (1975) |
| Hydroxymethylglutaryl CoA reductase (hamster) |                                               | Morgan et al. (1975) |
| Immunoglobulin D, delta (heavy) chain (human) |                                               | Morgan et al. (1975) |
| Immunoglobulin, heavy chain V-III region (human) |                                               | Morgan et al. (1975) |
| Immunoglobulin, x-light chain V-I region (human) |                                               | Morgan et al. (1975) |
| Immunoglobulin G, lambda (light) chain (human Sm) |                                               | Morgan et al. (1975) |
| Immunoglobulin G, lambda (light) chain (human NEI) |                                               | Morgan et al. (1975) |
| Immunoglobulin G1, heavy chain V-II region (human) |                                               | Morgan et al. (1975) |
| Inhibin beta A chain precursor (pig)           |                                               | Morgan et al. (1975) |
| Insulin receptor (human)                      |                                               | Morgan et al. (1975) |
| Interferon gamma (human)                      |                                               | Morgan et al. (1975) |
| Invertase (yeast)                             |                                               | Morgan et al. (1975) |
| J-chain (human)                               |                                               | Morgan et al. (1975) |
| Lectin (Dolichus biflorus seed)               |                                               | Morgan et al. (1975) |
| Leukocyte adhesion glycoprotein Mac-1 (human) |                                               | Morgan et al. (1975) |
| Link protein (rat)                            |                                               | Morgan et al. (1975) |
| Lymphotoxin (human)                           |                                               | Morgan et al. (1975) |
| Lysosomal membrane glycoprotein lamp-2 (human) |                                               | Morgan et al. (1975) |
| Lysosomal membrane glycoprotein lamp-A (human) |                                               | Morgan et al. (1975) |
| M1-1 protoxin (yeast)                         |                                               | Morgan et al. (1975) |
| Macrophage-lymphocyte Fc receptor (mouse)    |                                               | Morgan et al. (1975) |
| MRC OX-2 antigen (rat)                        |                                               | Morgan et al. (1975) |
Table I. Continued

| Proteins containing gs⁺ sites only | gs⁺ sites excluded due to homology | Reference |
|-----------------------------------|----------------------------------|-----------|
| Mylein PO protein (rat)           |                                  | NBRF      |
| Neutrophil elastase (human)       |                                  | Takahashi et al. (1988) |
| Nidogen (mouse)                   |                                  | Mann et al. (1989)      |
| Peroxidase (turnip)               |                                  | NBRF      |
| Phaseolin (Phaseolus vulgaris)    |                                  | Stitham et al. (1985)   |
| Pheromone prepro-alpha-factor (yeast) |                              | Winters et al. (1988) |
|                                   |                                  | En ter et al. (1983)    |
|                                   |                                  | Julius et al. (1984)    |
|                                   |                                  | Kusmin and Herskowitz (1982) |
| Phospholipase A2 (honeybee)       |                                  | NBRF      |
| Phospholipase A2 (Australian taipan) |                                | NBRF      |
| Plasma kallikrein (human)         |                                  | NBRF      |
| Plasminogen (bovine)              |                                  | NBRF      |
| Platelet-derived growth factor A chain (human) | | NBRF      |
| Platelet glycoprotein Ib alpha chain (human) | | NBRF      |
| Poly-immunoglobulin receptor (human) |                                | NBRF      |
| Postheparin plasma hepatic triglyceride lipase (human) | | NBRF      |
| Procalcitonin receptor (rat)      |                                  | Martin et al. (1988)    |
| Proopiomelanocortin (Xenopus laevis) |                                | Boutil et al. (1988)   |
| Prostatein (rat)                  |                                  | Maints (1986)           |
| Protective protein (human)        |                                  | Galja et al. (1988)     |
| Protein C light chain (bovine)    |                                  | Stenflo and Stenflo (1982) |
| Protein C heavy chain (bovine)    |                                  | NBRF      |
| Protein Z (bovine)                |                                  | Kraussus and Rauslaht (1986) |
| Proteoglycan core protein (human) |                                  | MacGillivray and Davie (1984) |
| Prothrombin (bovine)              |                                  | NBRF      |
| Rezinoblastoma-associated protein (human) |                            | NBRF      |
| Rhodopsin (bovine)                |                                  | Ovchinnikov (1982)      |
| Riboflavin-binding protein (chicken) |                              | Hargrave (1977)         |
| Ribonuclease, secretory (human urine) |                            | Harrazume el at. (1984) |
| Ribonuclease (porcine pancreas)   | 34⁺, 76⁺                        | Bentemra et al. (1988)  |
| Ricin D, B chain (castor bean)    | 21⁺, 3⁺                        | Jackson and Hirs (1970) |
| Stellacyanin (Japanese lacquertree) |                                | NBRF      |
| Structural glycoprotein E1 (SFV)  | 141⁺                           | Garoff et al. (1980)    |
| Structural glycoprotein E2 (SFV)  | 262⁺                           | Garoff et al. (1980)    |
| Structural glycoprotein E1 (Sindbis virus) |                  | Rice and Strauss (1981) |
| Structural glycoprotein E2 (Sindbis virus) |                     | Rice and Strauss (1981) |
| Structural glycoprotein E3 (Sindbis virus) |                  | Rice and Strauss (1981) |
| Sty lar glycoprotein 2 (winged tobacco) |                      | NBRF      |
| T-cell receptor alpha chain C region (mouse) |                  | NBRF      |
| T-cell receptor alpha chain V region (human) |                  | NBRF      |
| T-cell receptor beta chain V region (mouse) |                  | NBRF      |
| T-cell surface glycoprotein CD4 (human) |                  | NBRF      |
| T-cell surface glycoprotein CD8 (mouse) |                  | NBRF      |
| Thyroxine-binding globulin (human) |                                  | Fink et al. (1986)      |
| Tissue factor (human)             |                                  | Zinn et al. (1978)      |
| Transform receptor (human)        |                                  | Spicer et al. (1987)    |
| Transforming growth factor-alpha (human) |                    | Schneider et al. (1984) |
| Transmembrane protein E1 (coronavirus) |                   | Bringman et al. (1987)  |
| Urokinase-type plasminogen activator (human) |                 | Laude et al. (1987)     |
| Urokinase-type plasminogen activator (pig) |                  | NBRF      |
| Variant surface glycoprotein (Trypanosoma) |                | Bangs et al. (1988)     |
| [438](#)                                                                 |

1. According to experimental evidence presented in the reference.
2. According to experimental evidence cited in the reference.
3. In the reference, the absence of carbohydrate at this site is explicitly mentioned.
4. PTH-Asn was detected.
5. The relevant portion of the protein did not contain carbohydrate.
6. The Asn(180)-Gly(181) bond was susceptible to cleavage with hydroxylamine.
7. Among the other sites listed, there is at least one located in a sequence highly homologous to this one. Therefore, this site has not been included in the sequence statistics.
sequences there are additional potential glycosylation sites that could not be assigned or were discarded because they were located in transmembrane or cytoplasmic domains of the integral membrane proteins. These sites are not mentioned. Known partially glycosylated sites have been counted in the gs set. If a protein contains sites that had to be excluded from the statistics owing to homology, this is noted in the table.

Results

Non-standard sites

In a small number of cases, the sequence around the reported N-glycosylated asparagine did not agree with the Asn-X-Thr/Ser consensus. As was also noted by Nakai and Kanchisa (1988), three Asn-X-Cys patterns have been reported as gs sites in the NBRF database. The possibility of carbohydrate attachment at such sites was predicted by Bause and Leglier (1981). The Asn-X-Cys sites were found in bovine and human protein C and in human von Willebrand factor. However, we did not find any experimental evidence for glycosylation of the Asn-X-Cys sites in the reference given for human protein C (Foster et al., 1985).

We also found another unusual N-glycosylation site: in murine IgM heavy chain, carbohydrate is found bound to asparagine in the sequence Asn-Gly-Gly-Thr. A similar site has been reported for egg yolk phosvitin, a protein derived from vitellogenin. In this case, the sequence at the point of attachment was reported as Asn-Ser-Gly-Psr, where Psr is phosphoserine (Shainkin and Perlmann, 1971). However, the nucleotide sequence (van het Schip et al., 1987; Byrne et al., 1984) indicates that the site is of the normal Asn-Gly-Ser type.

Thus, although some of the putative non-standard sites may have been erroneously identified, at least a couple remain that seem to be authentic (Titani et al., 1986; Kehry et al., 1979; Stenflo and Fernlund, 1982). In exceptional cases, then, N-linked glycosylation does not seem to require the Asn-X-Ser/Thr consensus.

Statistical analysis

As can be seen in Table I, gs sites are far more common in glycoproteins than are gs sites. Apparently, if the oligosaccharyl transferase is present, the Asn-X-Thr/Ser signal leads to glycosylation approximately nine times out of 10.

In order to compare gs and gs sequences, we extracted 33-residue segments centred around the glycosylation sites listed in Table I. Amino acid distributions were calculated for gs and gs sites separately. The results for the residues immediately surrounding the consensus tripeptide are shown in Table II.

According to previous statistical studies, Pro is very rare or even absent in position +1 of gs sites (Mononen and Karjalainen, 1984). The statistical significance of this observation is confirmed by our data (Figure 1; P < 2 x 10^-5) as estimated from a binomial distribution with P = 0.0558, i.e. the mean frequency of Pro outside positions 0 to +3. Actually, the frequency of glycosylated Asn-Pro-Thr/Ser sites may be even lower since the Pro-containing site in thyroxine-binding globulin may have been erroneously identified as a gs site (an Asn-Cys-Thr acceptor site in position 233 seems to have been overlooked; cf. Flink et al., 1986; Zinn et al., 1978).

In addition to the under-representation of Pro in position +1 of the gs sites, there is another significant (P < 2 x 10^-5) frequency drop in position +3. In contrast, for the gs sites, Pro is significantly enriched in position +1 (P < 6 x 10^-5). The second most proline-rich position in gs sequences is +3, although the over-representation in this position is not statistically significant. The +3 pattern has not been noted previously, but

| Amino Acid | Position |
|------------|----------|
|            | -5 | -4 | -3 | -2 | -1 | 0  | +1 | +2 | +3 | +4 | +5 |
| Asn         | 5.8| 5.1| 5.5| 5.0| 6.2| 0.0| 8.4| 0.0| 7.2| 7.7| 6.7|
| Cys         | 3.4| 3.4| 3.4| 1.9| 1.7| 0.0| 2.6| 0.5| 4.3| 2.2| 3.8|
| Thr         | 4.4| 5.3| 3.4| 3.1| 3.8| 0.0| 4.3| 0.0| 4.8| 4.3| 5.5|
| Ser         | 4.6| 6.1| 4.6| 3.6| 3.5| 0.0| 2.9| 0.0| 6.5| 5.8| 4.3|
| Gly         | 4.1| 4.1| 3.6| 6.5| 4.8| 0.0| 5.0| 0.0| 3.6| 5.0| 4.1|
| Pro         | 6.8| 5.6| 9.2| 5.3| 8.6| 0.0| 11.0| 0.2| 5.8| 6.0| 7.2|
| Cys         | 2.7| 3.4| 2.7| 4.1| 1.7| 0.0| 2.4| 0.0| 2.2| 2.6| 2.4|
| Thr         | 4.4| 4.6| 5.3| 4.0| 2.0| 0.0| 5.8| 0.0| 5.5| 4.6| 5.3|
| Ser         | 4.4| 5.6| 5.3| 6.5| 3.8| 0.0| 4.8| 0.0| 3.6| 4.3| 5.0|
| Gly         | 11.4| 9.7| 9.6| 7.2| 8.9| 0.0| 8.4| 0.0| 9.6| 8.2| 9.1|
| Pro         | 4.4| 4.7| 1.7| 1.9| 2.2| 0.0| 1.7| 0.0| 2.2| 1.9| 1.4|
| Cys         | 3.9| 4.4| 2.7| 5.8| 4.8| 100.0| 4.8| 0.0| 3.8| 3.6| 4.1|
| Thr         | 4.8| 5.8| 6.5| 7.0| 5.8| 0.0| 5.0| 0.0| 1.9| 6.5| 6.7|
| Ser         | 3.6| 4.4| 5.5| 4.6| 4.3| 0.0| 4.8| 0.0| 4.3| 3.6| 4.8|
| Gly         | 4.6| 3.4| 5.1| 6.0| 4.3| 0.0| 4.8| 0.0| 4.3| 3.6| 4.8|
| Pro         | 9.0| 6.5| 7.2| 7.9| 9.8| 0.0| 9.8| 3.6| 8.9| 8.2| 6.2|
| Cys         | 6.5| 5.8| 5.8| 5.8| 6.7| 0.0| 6.2| 65.7| 5.3| 9.8| 6.2|
| Thr         | 5.6| 8.5| 5.8| 7.9| 7.7| 0.0| 9.1| 0.0| 10.3| 7.9| 7.9|
| Ser         | 1.5| 1.0| 2.4| 1.4| 1.9| 0.0| 1.0| 0.0| 1.4| 1.7| 1.7|
| Gly         | 3.9| 3.9| 5.1| 4.6| 5.5| 0.0| 4.3| 0.0| 4.6| 3.6| 3.8|

Frequencies are given as percentages. Position 0 is the carbohydrate attachment point.
is consistent with the findings of Bause (1983) and Roitsch and Lehle (1989) that peptides containing a potential glycosylation site cannot be glycosylated if the site has Pro in position +1 or +3.

Low counts in position +1 of gs+ sites have been reported for other amino acids besides Pro. For example, Cys and Trp (Kaplan et al., 1987; Lennarz, 1988) as well as the acidic residues, Glu and Asp (Mononen and Karjalainen, 1984) have been claimed to be rare in glycosylated tripeptides. Our data do not substantiate this, and there are no residues besides Pro for which a high frequency in some position close to gs~ sites is matched by a low frequency close to gs+ sites, or vice versa.

For proteins in general, i.e. including cytoplasmic ones, the frequencies of Asn-X-Thr and Asn-X-Ser tripeptides are equal. However, the sequence Asn-X-Thr has been reported to be about three times as frequent as Asn-X-Ser for gs+ sites (Kaplan et al., 1987; Lennarz, 1988). In agreement with this, experiments by Bause (1984) indicate that the tripeptide Asn-X-Thr is glycosylated more rapidly than Asn-X-Ser. Our data essentially confirm these results.

Fig. 1. Frequency of proline residues as a function of position relative to glycosylated (A) and non-glycosylated (B) Asn-X-Thr/Ser sites. The Asn residue is in position zero.

In order to look for correlations between glycosylation tendency and position in the protein, we have tabulated the relative positions of gs+ and gs~ sites in the sequence and the absolute distances to nearest gs+ sites as well as to the N and C termini of the protein. From these data, we find that the frequency of gs~ sites is higher towards the C terminus, whereas that of gs+ sites is lower. The distributions (Figure 2) differ significantly ($P < 10^{-2}$ by $x^2$ test).

In terms of absolute distances, we have found one gs+ site only four residues away from the C terminus of a mature glycoprotein and one only one residue away from the N terminus.

We have also calculated distributions for the distances between neighbouring gs+ sites. The smallest separation in our database is four residues (in haptoglobin-1 with the sequence Asn-His-Ser-Glu-Asn-Ala-Thr). It may be that steric hindrance prevents more closely spaced sites from being glycosylated at the same time; thus, a site with the sequence . . . Asn-Asn-Ser-Thr . . . with the second but not the first Asn glycosylated is found in human cholinesterase (see Table I).

Fig. 2. Incidence of glycosylated (open squares) and non-glycosylated (solid squares) Asn-X-Thr/Ser sites as a function of relative position in the protein chain (N- to C-terminal direction).

Fig. 3. Mean turn propensities as a function of position relative to glycosylated (A) and non-glycosylated (B) Asn-X-Thr/Ser sites.

Beeley (1977) as well as Aubert et al. (1976) have noted that gs+ sites are often situated in beta turns or other loop structures. We have thus calculated the mean turn, $\beta$-structure, and $\alpha$-helix potentials as a function of position using the scale of Levitt (1978). We find no indications that gs+ sites are more turn-prone than gs~ sites (Figure 3), confirming earlier results (Mononen and
Karjalainen, 1984). Since both asparagine and the hydroxy amino acids have fairly high turn propensities, turn probability should not be expected to differ much between gs+ and gs sites. Actually, the mean turn potential of gs sites is a little higher than that of gs sites because of the high frequency of proline residues. Thus, the importance of turn conformations for glycosylation remains conjectural.

Conclusions

Our data demonstrate that potential glycosylation sites that are located near the C terminus or contain Pro in position +1 or +3 often are non-glycosylated. These observations allow a clear prediction of the likelihood for glycosylation to be made in some cases, but not in others. A site with Pro is non-glycosylated in at least 90% of all cases. For Pro, this value is ~50%. Thus, one can safely assume a site with Pro to be gs and a site lacking Pro to be gs+ (correct 93% of the time). Sites with Pro are best left undecided. Thus, out of a total of 465 sites, 10/11 sites (90%) with Pro are correctly predicted as gs+, 369/400 sites (92%) not having Pro are correctly predicted as gs+, and 6/465 sites with Pro (1%) cannot be predicted. Beyond this, no simple local patterns have been found to correlate with glycosylation; in particular, secondary structure predictions are of no use in differentiating between gs+ and gs- sites.

Experiments with ovalbumin have shown that co-translational glycosylation can take place until ~45 amino acids have been added beyond the carbohydrate attachment site (Glabe et al., 1980). Therefore, the bias of gs sites away from the C terminus may indicate that glycosylation occurs more easily when the nascent polypeptide chain is still attached to polyribosomes and spanning the ER membrane. N-Terminally located sites spend more time in the neighbourhood of the lumenal face of the ER membrane, where the sugar-adding enzyme is situated. Also, C-terminal sites may be more quickly pulled into an already almost fully folded structure, thus becoming inaccessible for glycosylation.

In conclusion, we find that gs sites tend to be found more frequently towards the C terminus of glycoproteins, and that proline residues in positions X and Y in the consensus Asn-X-Thr/Ser-Y +3 often are non-glycosylated. These observations allow a clear prediction of the likelihood for glycosylation to be made in some cases, but not in others. A site with Pro is non-glycosylated in at least 90% of all cases. For Pro, this value is ~50%.

Acknowledgements

This work was supported by a grant from the Swedish Natural Sciences Research Council to G.vH.

References

Aubert, J.-P., Biserte, G. and Loucheux-Lefebvre, M.-H. (1976) Arch. Biochem. Biophys., 175, 410-418.
Bangs, J.D., Doering, T.L., Englund, P.T. and Hart, G.W. (1988) J. Biol. Chem., 263, 17697-17705.
Baudys, M. and Kostka, V. (1983) Eur. J. Biochem., 123, 121-131.
Beeley, J.G. (1977) Biochim. Biophys. Acta, 492, 83-94.
Beeley, J.G., Biserte, G. and Loucheux-Lefebvre, M.-H. (1976) Biochim. Biophys. Acta, 492, 83-94.
Beeley, J.G., Biserte, G. and Loucheux-Lefebvre, M.-H. (1976) Biochim. Biophys. Acta, 492, 83-94.
Beeley, J.G., Biserte, G. and Loucheux-Lefebvre, M.-H. (1976) Biochim. Biophys. Acta, 492, 83-94.
Beeley, J.G., Biserte, G. and Loucheux-Lefebvre, M.-H. (1976) Biochim. Biophys. Acta, 492, 83-94.
Beeley, J.G., Biserte, G. and Loucheux-Lefebvre, M.-H. (1976) Biochim. Biophys. Acta, 492, 83-94.
Beeley, J.G., Biserte, G. and Loucheux-Lefebvre, M.-H. (1976) Biochim. Biophys. Acta, 492, 83-94.
