Glycoconjugate Journal 15, 507–509 (1998)

EGC2
Glycosyltransferases in SWISS-PROT

Claire O'Donovan* and Nicoletta Mitaritonna‡

The EMBL Outstation – The European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

SWISS-PROT is a curated protein sequence database with a high level of annotation (such as description of the function of a protein, its domain structure, post-translational modification, variants, etc), a minimal level of redundancy and a high level of integration with other databases. An ongoing project is to maintain the glycosyltransferase family of enzymes with comprehensive annotation and documentation in the SWISS-PROT database and to represent the most recent research developments.

Introduction

SWISS-PROT was established in 1986 and is maintained collaboratively by the EMBL Outstation – The European Bioinformatics Institute (EBI) [1] and the Department of Medical Biochemistry at the University of Geneva. The data in SWISS-PROT are derived mainly from translation of DNA sequences from the EMBL Nucleotide Sequence Database, submitted directly by researchers or extracted from the literature. SWISS-PROT is committed to providing a valuable resource for researchers with a comprehensive current representation of protein sequences and information. As part of this, we select particular projects for special attention. These projects cover a wide range of topics such as protein families and model organisms and we make use of external experts, who make time to send us their comments and updates concerning the specific topics. It is within this context that our work on glycosyltransferases takes place.

SWISS-PROT and glycosyltransferases

Our aims for this glycosyltransferase project are:

1. To maintain this enzyme family with fully comprehensive annotation and documentation in the SWISS-PROT database.

All known glycosyltransferase sequences are present either in SWISS-PROT or in its new computer-annotated supplement, TREMBL [2]. For each SWISS-PROT sequence entry, the core data consist of the sequence data, the citation information (bibliographical references) and the taxonomic data (description of the biological source of the protein), while the annotation consists of the description of the following items:

- Function(s) of the protein
- Post-translational modification(s). For example carbohydrates, phosphorylation, acetylation, GPI-anchor, etc.
- Domains and sites. For example calcium-binding regions, ATP-binding sites, zinc fingers, homeobox, kringle, etc.
- Secondary structure
- Quaternary structure
- Similarities to other proteins
- Disease(s) associated with deficiencies in the protein
- Sequence conflicts, variants, etc.

To achieve this annotation we use, in addition to the publications that report new sequence data, review articles to update the annotation of families or groups of proteins. We also use other relevant databases such as ENZYME and PROSITE to add alternative names, catalytic activities, cofactors and function information [3, 4]. An example of a glycosyltranferase entry is illustrated in Figure 1.

2. To represent the most recent research developments.

To achieve this goal, we are working closely with our external expert for glycosyltransferases, Iain Wilson, Carbohydrate Research Centre, University of Dundee, Scotland, who provides A GUIDE TO CLONED GLYCOSYLTRANSFERASES which he crosslinks to SWISS-PROT. By receiving his comments and updates concerning glycosyltransferases, we are kept firmly in touch with the research community and the developments therein.

At present, one of the areas of particular interest in the glycosyltransferase project which combines both the

*E-mail: odonovan@ebi.ac.uk
‡E-mail: nico@ebi.ac.uk

```
ID   BGAT_HUMAN     STANDARD;      PRT;   354 AA.
AC   P16442;
DT   01-AUG-1990 (REL. 15, CREATED)
DT   01-DEC-1992 (REL. 24, LAST SEQUENCE UPDATE)
DT   01-NOV-1995 (REL. 32, LAST ANNOTATION UPDATE)
DE   FUCOSYLGLYCOPROTEIN ALPHA-N-ACETYLGALACTOSAMINYLTRANSFERASE
DE   (EC 2.4.1.40) (HISTO-BLOOD GROUP A TRANSFERASE) (A TRANSFERASE) /
DE   FUCOSYLGLYCOPROTEIN 3-ALPHA-GALACTOSYLTRANSFERASE (EC 2.4.1.37)
DE   (HISTO-BLOOD GROUP B TRANSFERASE) (B TRANSFERASE) (NAGAT).
GN   ABO.
OS   HOMO SAPIENS (HUMAN).
OC   EUKARYOTA; METAZOA; CHORDATA; VERTEBRATA; TETRAPODA; MAMMALIA;
OC   EUTHERIA; PRIMATES.
RN   [1]
RP   SEQUENCE FROM N.A., AND PARTIAL SEQUENCE.
RX   MEDLINE; 90110098.
RA   YAMAMOTO F.-I., MARKEN J., TSUJI T., WHITE T., CLAUSEN H.,
RA   HAKOMORI S.-I.;
RL   J. BIOL. CHEM. 265:1146-1151(1990).
RN   [2]
RP   SEQUENCE FROM N.A.
RX   MEDLINE; 90238543.
RA   YAMAMOTO F.-I., CLAUSEN H., WHITE T., MARKEN J., HAKOMORI S.-I.;
RL   NATURE 345:229-233(1990).
RN   [3]
RP   CHARACTERIZATION.
RX   MEDLINE; 91035461.
RA   YAMAMOTO F.-I., HAKOMORI S.-I.;
RL   J. BIOL. CHEM. 265:19257-19262(1990).
CC   -!- FUNCTION: THIS PROTEIN IS THE BASIS OF THE ABO BLOOD GROUP SYSTEM.
CC       THE HISTO-BLOOD GROUP ABO INVOLVES THREE CARBOHYDRATE ANTIGENS: A,
CC       B, AND H. A, B, AND AB INDIVIDUALS EXPRESS A GLYCOSYLTRANSFERASE
CC       ACTIVITY THAT CONVERTS THE H ANTIGEN TO THE A ANTIGEN (BY ADDITION
CC       OF UDP-GALNAC) OR TO THE B ANTIGEN (BY ADDITION OF UDP-GAL),
CC       WHEREAS O INDIVIDUALS LACK SUCH ACTIVITY. THE O PHENOTYPE IS
CC       RESULT OF A SINGLE BASE FRAMESHIFT DELETION IN THE N-TERMINAL
CC       EXTREMITY OF THE GENE.
CC   -!- CATALYTIC ACTIVITY: UDP-N-ACETYL-D-GALACTOSAMINE + GLYCOPROTEIN
CC       ALPHA-L-FUCOSYL-(1,2)-D-GALACTOSE = UDP + N-ACETYL-ALPHA-D-
CC       GALACTOSAMINYL-(1,3)-[ALPHA-L-FUCOSYL-(1,2)]-D-GALACTOSE.
CC   -!- CATALYTIC ACTIVITY: UDP-GALACTOSE + GLYCOPROTEIN ALPHA-L-FUCOSYL-
CC       (1,2)-D-GALACTOSE = UDP + GLYCOPROTEIN ALPHA-D-GALACTOSYL-(1,3)-
CC       [ALPHA-L-FUCOSYL-(1,2)]-D-GALACTOSE.
CC   -!- PATHWAY: GLYCOSYLATION.
CC   -!- SUBCELLULAR LOCATION: TYPE II MEMBRANE PROTEIN. MEMBRANE-BOUND
CC       FORM IN TRANS CISTERNAE OF GOLGI. SOLUBLE FORM IN BODY FLUIDS.
CC   -!- PTM: THE SOLUBLE FORM DERIVES FROM THE MEMBRANE FORM BY
CC       PROTEOLYTIC PROCESSING.
CC   -!- SIMILARITY: STRONG, TO N-ACETYLLACTOSAMINIDE ALPHA-1,3-
CC   -!- SIMILARITY: STRUCTURAL SIMILARITY WITH THE OTHER MAMMALIAN
CC       GLYCOSYLTRANSFERASES.
CC   -!- POLYMORPHISM: THE SEQUENCE SHOWN IS THAT OF THE A TRANSFERASE.
DR   EMBL; J05175; G340078; -.
DR   PIR; A34933; A34933.
DR   PIR; S09593; S09593.
DR   MIM; 110300;  11TH EDITION.
KW   TRANSFERASE; GLYCOSYLTRANSFERASE; GLYCOPROTEIN; TRANSMEMBRANE;
KW   SIGNAL-ANCHOR; GOLGI STACK; POLYMORPHISM.
FT   DOMAIN        1     32       CYTOPLASMIC (POTENTIAL).
FT   TRANSMEM     33     53       SIGNAL-ANCHOR (TYPE-II MEMBRANE PROTEIN).
FT   DOMAIN       54    354       LUMENAL, CATALYTIC (POTENTIAL).
FT   CHAIN        54    354       FUCOSYLGLYCOPROTEIN ALPHA-N-
FT                                ACETYLGALACTOSAMINYLTRANSFERASE, SOLUBLE
FT                                FORM.
```

Figure 1. A glycosyltransferase entry.

```
FT   CARBOHYD    113     113       POTENTIAL.
FT   VARIANT      80      81       CR -> W.
FT   VARIANT     156     156       P -> L.
FT   VARIANT     176     176       R -> G (IN GROUP B TRANSFERASE).
FT   VARIANT     235     235       G -> S (IN GROUP B TRANSFERASE).
FT   VARIANT     266     266       L -> M (IN GROUP B TRANSFERASE;
FT                                 (IMPORTANT FOR THE SPECIFICITY).
FT   VARIANT     268     268       G -> A (IN GROUP B TRANSFERASE;
FT                                 (IMPORTANT FOR THE SPECIFICITY).
SQ   SEQUENCE   354 AA;  40934 MW;  9F190385 CRC32;
     MAEVLRTLAG KPKCHALRPM ILFLIMLVLV LFGYGVLSPR SLMPGSLERG FCMAVREPDH
     LQRVSLPRMV YPQPKVLTPC RKDVLVVTPW LAPIVWEGTF NIDILNEQFR LQNTTIGLTV
     FAIKKYVAFL KLFLETAEKH FMVGHRVHYY VFTDQPAAVP RVTLGTGRQL SVLEVRAYKR
     WQDVSMRRME MISDFCERRF LSEVDYLVCV DVDMEFRDHV GVEILTPLFG TLHPGFYGSS
     REAFTYERRP QSQAYIPKDE GDFYYLGGFF GGSVQEVQRL TRACHQAMMV DQANGIEAVW
     HDESHLNKYL LRHKPTKVLS PEYLWDQQLL GWPAVLRKLR FTAVPKNHQA VRNP
```

Figure 1. Continued.

annotation and research aspects, is the organization of the glycosyltransferases into subfamilies. Grouping glycosyltransferases according to EC numbers is not sufficiently exclusive as within a particular EC number, the functions and the affinities of the proteins vary. They need to be represented in such a way as to enable proper annotation and meaningful comparisons of the proteins. In addition, there are glycosyltranferases which have yet been assigned EC numbers. We believe that SWISS-PROT is in an ideal position to play a role due to the centralization of information and its close collaboration with the ENZYME database. Currently, we are incorporating all glycosyltransferase sequences into SWISS-PROT and simultaneously, collating catalytic information for the assignment of EC numbers.

References

1 Rodriguez-Tome P, Stoehr PJ, Cameron GN, Flores TP (1996) *Nucleic Acids Res* **24**: 6–13.
2 Bairoch A, Apweiler R (1996) *Nucleic Acids Res* **24**: 21–5.
3 Bairoch A (1996) *Nucleic Acids Res* **24**: 221–2.
4 Bairoch A, Bucher P, Hofmann K (1996) *Nucleic Acids Res* **24**: 189–96.

Accepted December 1996