Identification of linkage-specific sequence motifs in sialyltransferases

Ronak Y. Patel and Petety V. Balaji1

School of Biosciences and Bioengineering, Indian Institute of Technology Bombay, Powai, Mumbai 400 076, India

Received on June 24, 2005; revised on September 26, 2005; accepted on September 28, 2005

Eukaryotic sialyltransferases (SiaTs) comprise a superfamily of enzymes catalyzing the transfer of sialic acid (Sia) from a common donor substrate to various acceptor substrates in different linkages. These enzymes have been classified as ST3Gal, ST6Gal, ST6GalNAc, and ST8Sia families based on linkage- and acceptor monosaccharide-specificities and sequence similarities. It was recognized early on that SiaTs contain certain well-conserved motifs, and these were denoted as L (large), S (small), and VS (very small)-motifs; recently, a fourth motif, denoted as motif III, was identified. These four motifs are common to all the SiaTs, irrespective of the linkage- and acceptor saccharide-specificities. In this study, the sequences of the various families have been analyzed, and sequence motifs that are unique to the various families have been identified. These unique motifs are expected to contribute to the characteristic linkage- and acceptor saccharide-specificities of the family members. One of the linkage-specific motifs is contiguous to L-motif. Members of ST3Gal and ST8Sia families share significant sequence similarities; in contrast, the ST6Gal family is distinct from the ST6GalNAc family. The latter consists of two subfamilies, one comprising ST6GalNAc I and ST6GalNAc II, and the other comprising ST6GalNAc III, ST6GalNAc IV, ST6GalNAc V, and ST6GalNAc VI. Each of these subfamilies has characteristic sequence motifs not present in the other subfamily.

Key words: acceptor specificity/linkage specificity/profile HMM

Introduction

Eukaryotic sialyltransferases (SiaTs) participate in glycan biosynthesis and catalyze the transfer of sialic acid (Sia) from cytidine monophosphate-Sia (CMP-Sia) donor substrate to acceptor glycolipid/glycoprotein. Sia can be found α2→3 linked to Gal-R, α2→6 linked to Gal-R, GalNAc-R, or GlcNAc-R and α2→8/9 linked to Sia-R; here, -R denotes the rest of the acceptor substrate moiety (Tsujii, 1996; Harduin-Lepers et al., 2001). SiaTs are classified as ST3Gal, ST6Gal, ST6GalNAc, and ST8Sia families based on the linkage in which Sia is transferred and the acceptor saccharide.

It was recognized early on that SiaTs contain certain conserved sequence motifs; these were denoted as L- (large), S- (small), and VS- (very small) motifs (Gillespie et al., 1992; Wen et al., 1992; Drickamer, 1993; Livingston and Paulson, 1993; Datta and Paulson, 1997; Geremia et al., 1997). In fact, this conserved sequence motifs were instrumental in the identification and cloning of additional SiaTs (e.g., Livingston and Paulson, 1993; Kitagawa and Paulson, 1994; Lee et al., 1994; Harduin-Lepers et al., 2000). These three motifs are conserved in all the SiaTs. Recently, an additional motif, denoted as motif III, was found to be conserved across all the SiaTs (Jeanneau et al., 2004). The sequence motifs/residues that are conserved across all the SiaTs may have an important role in the folding and/or maintenance of the three-dimensional structure which is expected to be common for all the SiaTs. Alternatively, the conserved motifs/residues may take part in functional aspects that are common to all the SiaTs, such as donor substrate binding and the various stages of catalysis. In fact, site-directed mutagenesis of rat ST6Gal I showed that the residues in the L-motif are involved in donor substrate binding (Datta and Paulson, 1995), and those in the S-motif are involved in donor as well as acceptor substrate binding (Datta et al., 1998). In human ST3Gal I, mutation of the conserved His or Tyr present in motif III to Ala and similarly, mutation of the conserved His present in the VS-motif to Ala resulted in loss of activity (Jeanneau et al., 2004). In ST8Sia II and ST8Sia IV also, mutation of the conserved His of the VS-motif to Lys resulted in loss of activity; it was also noted that this mutation did not affect folding or substrate binding (Kitazume-Kawaguchi et al., 2001).

The ST3Gal, ST6Gal, ST6GalNAc, and ST8Sia families differ from each other in their linkage specificity and in the nature of the saccharide moiety that accepts the Sia. SiaTs of ST3Gal and ST6Gal families transfer Sia to Gal-R, whereas those of ST8Sia family transfer to Sia-R. As a consequence of this, the binding pocket which accommodates the acceptor saccharide moiety in ST3Gal and ST6Gal families is expected to be different from that in the ST8Sia family. In addition, the binding pockets in ST3Gal and ST6Gal families will have to accommodate the same saccharide (i.e., galactose) in different orientations relative to the donor substrate (CMP-Sia) to enable the transfer of Sia in different linkages namely, α2→3 or α2→6. Subtle differences, necessary to accommodate the Gal or GalNAc, should also exist between the ST6Gal and ST6GalNAc families.

This sequence analysis study was undertaken to identify the sequence motifs that are specific to the ST3Gal, ST8Sia, ST6Gal, and ST6GalNAc families. The results obtained show that such linkage-(family-) specific sequence motifs do exist. These sequence motifs presumably contribute to
differences in the linkage- and saccharide-specificity of the different families. The newly identified family-specific sequence motifs will help not only in understanding the structure–function relationship of these enzymes but also in inferring the linkage specificity of putative SiaTs, identified from whole-genome sequencing studies.

Methods

Databases and software

The sequences of the experimentally characterized SiaTs were retrieved from the UniProt (Bairoch et al., 2005) and NCBI (www.ncbi.nlm.nih.gov) protein databases. The sequences of the SiaTs, computationally annotated based on their sequence similarity to the experimentally characterized SiaTs, were taken from the carbohydrate active enzyme (CAZy) database (Coutinho and Henrissat, 1999). Multiple sequence alignments were performed locally using Tcoffee (version 2.03; Notredame et al., 2000). Alignments were visualized using BioEdit (Hall, 1999), hmmbuild and hmmevalibrate modules of HMMER (version 2.3.2; Eddy, 1998) were used to generate hidden Markov model (HMM) profiles; the hmmssearch module was used for searching the databases using these profiles. Sequence logos were created using WebLogo (version 2.8.1; Crooks et al., 2004). All the software were used with default parameters, unless specified.

Creation of data sets and analysis strategy

The SiaTs were divided based on their linkage specificity and/or acceptor specificity into four classes, ST3Gal, ST8Sia, ST6Gal, and ST6GalNAc. Two data sets were created: one included 47 SiaTs (14 ST3Gal + 13 ST8Sia + 8 ST6Gal + 12 ST6GalNAc; Table I) whose enzyme activity and linkage specificity have been experimentally demonstrated; these were identified by literature search, and their amino acid sequences were retrieved from the UniProt or NCBI protein database. The second data set included 147 SiaTs (53 ST3Gal + 47 ST8Sia + 16 ST6Gal + 31 ST6GalNAc) which have been computationally annotated based on their sequence similarity to the experimentally characterized SiaTs. These are members of the family 29 of CAZY database which includes only SiaTs (EC 2.4.99–). Certain SiaTs in the CAZY database have more than one entry; in such cases, the latest entry was chosen.

Analyses were performed (1) by considering only the experimentally characterized SiaTs and (2) by including even the computationally annotated SiaTs. Although the larger data set size in the latter case enhances the statistical significance, inferences have been drawn only based on the analysis of experimentally characterized SiaTs. This has been done so, because, despite a high overall sequence similarity, changes in key residues may confer a different substrate/linkage specificity to enzymes. For example, the gene that was initially identified as β3GalT-3 based on sequence similarity (Amado et al., 1998) was subsequently shown to be β3GalNAcT (Okajima et al., 2000a).

Analyses were performed by putting together the SiaTs of all the four families namely, ST3Gal, ST8Sia, ST6Gal, and ST6GalNAc and by considering each family separately. The former provides information about the residues/motifs that are common to the SiaT family, whereas the latter provides information about the residues/motifs that are linkage- (family-) specific.

The conserved regions obtained from the various multiple sequence alignments were used for generating sequence logos. Sequence logos are better than consensus sequences to describe the conserved regions (Schneider and Stephens, 1990; Mount, 2003; http://weblogo.berkeley.edu). A sequence logo shows the relative frequencies of the various residues at a given position by proportionally varying the size of the symbol. The order of predominance of the residues at a given position are indicated by showing the most (or least) frequently occurring residue at the top (or bottom) of the heap. The height of the logo at a given position is proportional to the degree of conservation or the importance/information content of that position.

Results

Variations are observed in the L-, S-, and VS-motifs and motif III across the families

The sequences of all the 47 experimentally characterized SiaTs (Table I) were multiply aligned. In this alignment, the previously reported (Harduin-Lepers et al., 2001; Krzewinski-Recchi et al., 2003; Jeanneau et al., 2004) sequence motifs conserved in all the SiaTs viz., L-, S-, and VS-motifs and motif III are all discernible as expected (Figure S1). Sequence logos, generated using the alignment corresponding to these four motifs, clearly show the residues that are completely conserved in all the SiaTs (Figure 1).

Identification of linkage-specific sequence motifs

In addition to the L-, S-, and VS-motifs and motif III reported earlier, other, linkage- (family-) specific conserved motifs were identified from the multiple sequence alignment of experimentally characterized SiaTs (Figure 2). Such linkage-specific motifs in the ST3Gal, ST8Sia, and ST6Gal families are evident from their respective multiple sequence alignments. These motifs are also present in the computationally annotated SiaTs belonging to the respective families. These motifs are small, consisting of six to about twenty residues. One of the motifs in each of the three families immediately follows the L-motif (Figure S1). This motif has previously been identified in the ST3Gal (Okajima et al., 1999a) and ST8Sia (Angata and Fukuda, 2003) families and has been suggested to play a role in acceptor binding (Fukumoto et al., 1999). All the motifs, except the last ones in the ST6Gal family and ST6GalNAc subfamily I, are located between the L- and S-motifs.

In a sequence logo, the residues are depicted in such a way that their height is proportional to their frequency of occurrence at that position (Schneider and Stephens, 1990; Mount, 2003). Comparison of the various linkage-specific motifs shows that the fully conserved residues are of different heights, for example, conserved residues in the ST3Gal motifs are much taller than those in the ST6GalNAc subfamilies. This difference arises because of the differences in the number of sequences taken for multiple sequence alignment.
Table I. Length, accession number gene name, and common name of experimentally characterized sialyltransferases analyzed in this study

<table>
<thead>
<tr>
<th>Length</th>
<th>UniProt ID</th>
<th>Gene name</th>
<th>Common name</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ST3Gal family: 14 sequences, 6 nonorthologous genes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>340</td>
<td>Q11201</td>
<td>HSIAT4A</td>
<td>HST3Gal I</td>
</tr>
<tr>
<td>337</td>
<td>P54751</td>
<td>MSIAT4A</td>
<td>MST3Gal I</td>
</tr>
<tr>
<td>342</td>
<td>Q11200</td>
<td>CSIAT4A</td>
<td>CST3Gal I</td>
</tr>
<tr>
<td>343</td>
<td>Q02745</td>
<td>PSIAT4A</td>
<td>PST3Gal I</td>
</tr>
<tr>
<td>350</td>
<td>Q16842</td>
<td>HSIAT4B</td>
<td>HST3Gal II</td>
</tr>
<tr>
<td>349</td>
<td>NP_835149F</td>
<td>MSIAT4B</td>
<td>MST3Gal II</td>
</tr>
<tr>
<td>375</td>
<td>Q11203</td>
<td>HSIAT6</td>
<td>HST3Gal III</td>
</tr>
<tr>
<td>374</td>
<td>P97325</td>
<td>MSIAT6</td>
<td>MST3Gal III</td>
</tr>
<tr>
<td>374</td>
<td>Q02734</td>
<td>RSIAT6</td>
<td>RST3Gal III</td>
</tr>
<tr>
<td>333</td>
<td>Q11206</td>
<td>HSIAT4C</td>
<td>HST3Gal IV</td>
</tr>
<tr>
<td>333</td>
<td>NP_033204F</td>
<td>MSIAT4C</td>
<td>MST3Gal IV</td>
</tr>
<tr>
<td>362</td>
<td>Q9UNP4</td>
<td>HSIAT9</td>
<td>HST3Gal V</td>
</tr>
<tr>
<td>359</td>
<td>O88829</td>
<td>MSIAT9</td>
<td>MST3Gal V</td>
</tr>
<tr>
<td>331</td>
<td>Q9Y274</td>
<td>HSIAT10</td>
<td>HST3Gal VI</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ST8Sia family: 13 sequences, 6 nonorthologous genes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>356</td>
<td>Q92185</td>
<td>HSIAT8A</td>
<td>HST8Sia I</td>
</tr>
<tr>
<td>355</td>
<td>Q64687</td>
<td>MSIAT8A</td>
<td>MST8Sia I</td>
</tr>
<tr>
<td>375</td>
<td>Q92186</td>
<td>HSIAT8B</td>
<td>HST8Sia II</td>
</tr>
<tr>
<td>375</td>
<td>O35696</td>
<td>MSIAT8B</td>
<td>MST8Sia II</td>
</tr>
<tr>
<td>375</td>
<td>Q07977</td>
<td>RSIAT8B</td>
<td>RST8Sia II</td>
</tr>
<tr>
<td>380</td>
<td>O43173</td>
<td>HSIAT8C</td>
<td>HST8Sia III</td>
</tr>
<tr>
<td>380</td>
<td>Q64689</td>
<td>MSIAT8C</td>
<td>MST8Sia III</td>
</tr>
<tr>
<td>398</td>
<td>Q92187</td>
<td>HSIAT8D</td>
<td>HST8Sia IV</td>
</tr>
<tr>
<td>359</td>
<td>Q64692</td>
<td>MSIAT8D</td>
<td>MST8Sia IV</td>
</tr>
<tr>
<td>359</td>
<td>Q64690</td>
<td>CGIAT8D</td>
<td>CGST8Sia IV</td>
</tr>
<tr>
<td>376</td>
<td>O15466</td>
<td>HSIAT8E</td>
<td>HST8Sia V</td>
</tr>
<tr>
<td>412</td>
<td>P70126</td>
<td>MSIAT8E</td>
<td>MST8Sia V</td>
</tr>
<tr>
<td>398</td>
<td>Q8K4T1</td>
<td>MSIAT8F</td>
<td>MST8Sia VI</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ST6Gal family: 8 sequences, 3 nonorthologous genes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>406</td>
<td>P15907</td>
<td>HST6GALI</td>
<td>HST6Gal I</td>
</tr>
<tr>
<td>403</td>
<td>Q64685</td>
<td>MST6GALI</td>
<td>MST6Gal I</td>
</tr>
<tr>
<td>413</td>
<td>Q92182</td>
<td>CST6GALI</td>
<td>CST6Gal I</td>
</tr>
<tr>
<td>403</td>
<td>P13721</td>
<td>RST6GALI</td>
<td>RST6Gal I</td>
</tr>
<tr>
<td>405</td>
<td>O18974</td>
<td>BST6GALI</td>
<td>BST6Gal I</td>
</tr>
<tr>
<td>529</td>
<td>Q81UG7</td>
<td>HST6GALII</td>
<td>HST6Gal II</td>
</tr>
<tr>
<td>524</td>
<td>Q76K27</td>
<td>MST6GALII</td>
<td>MST6Gal II</td>
</tr>
<tr>
<td>474</td>
<td>Q9GU23</td>
<td>DST6GAL</td>
<td>DST6Gal</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ST6GalNAc family: 12 sequences, 6 nonorthologous genes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>600</td>
<td>Q9NSC7</td>
<td>HSIAT7A</td>
<td>HST6GalNAc I</td>
</tr>
<tr>
<td>526</td>
<td>Q9QZ39</td>
<td>MSIAT7A</td>
<td>MST6GalNAc I</td>
</tr>
<tr>
<td>566</td>
<td>Q92183</td>
<td>CSIAT7A</td>
<td>CST6GalNAc I</td>
</tr>
<tr>
<td>374</td>
<td>Q9U137</td>
<td>HSIAT7B</td>
<td>HST6GalNAc II</td>
</tr>
<tr>
<td>373</td>
<td>P70277</td>
<td>MSIAT7B</td>
<td>MST6GalNAc II</td>
</tr>
<tr>
<td>404</td>
<td>Q92184</td>
<td>CSIAT7B</td>
<td>CST6GalNAc II</td>
</tr>
<tr>
<td>305</td>
<td>Q9WUVC2</td>
<td>MSIAT7C</td>
<td>MST6GalNAc III</td>
</tr>
<tr>
<td>305</td>
<td>Q64686</td>
<td>RSIAT7C</td>
<td>RST6GalNAc III</td>
</tr>
</tbody>
</table>
Table I. continued

<table>
<thead>
<tr>
<th>Length</th>
<th>UniProt ID</th>
<th>Gene name&lt;sup&gt;ab&lt;/sup&gt;</th>
<th>Common name</th>
</tr>
</thead>
<tbody>
<tr>
<td>302</td>
<td>Q9H4F1</td>
<td>HSIA7D</td>
<td>HST6GalNAc IV</td>
</tr>
<tr>
<td>360</td>
<td>Q9R2B6</td>
<td>MSIA7D</td>
<td>MST6GalNAc IV</td>
</tr>
<tr>
<td>336</td>
<td>Q9QYJ1</td>
<td>MSIA7E</td>
<td>MST6GalNAc V</td>
</tr>
<tr>
<td>333</td>
<td>NP_058669&lt;sup&gt;c&lt;/sup&gt;</td>
<td>MSIA7F</td>
<td>MST6GalNAc VI</td>
</tr>
</tbody>
</table>

<sup>a</sup>The gene name is from the NCBI’s Entrez Gene database; this database also lists other aliases of the respective genes. The same name is also listed in the UniProt database.

<sup>b</sup>The prefix for the gene name indicates the name of the source organism: B, cow (Bos taurus); C, chicken (Gallus gallus); Cg, Chinese hamster (Cricetulus griseus); D, Drosophila melanogaster; H, human (Homo sapiens); M, mouse (Mus musculus); P, pig (Sus scrofa); R, rat (Rattus norvegicus).

<sup>c</sup>NCBI accession number.

Fig. 1. Sequence logos of the L-motif (top panel), S-motif (middle panel), motif III (bottom panel, left), and VS-motif (bottom panel, right). The conserved regions for generating the logos were extracted from the multiple sequence alignment of the 47 experimentally characterized SiaTs (Table I). The numbers along the abscissa indicate the position of residues within the conserved region. The ordinates are in units of bits and are indicative of the information content at each position (Schneider and Stephens, 1990; Mount, 2003).
Fig. 2. Sequence logos of the linkage- (family-) specific motifs. The schematic below the motifs indicates the relative positions of these motifs in the sequence and in the context of other, hitherto characterized motifs. Besides these motifs that are conserved either at the superfamily or at the family level, there are other residues that are conserved at the subfamily level (Harduin-Lepers et al., 2005). ST6GalNAc subfamily I consists of ST6GalNAc I and ST6GalNAc II, and ST6GalNAc subfamily II consists of ST6GalNAc III, ST6GalNAc IV, ST6GalNAc V, and ST6GalNAc VI. The conserved regions for generating the logos were extracted from the family-specific multiple sequence alignments (Figure S1) of the experimentally characterized SiaTs (Table I). The numbers along the abscissa indicate the position of residues within the conserved region. The ordinates are in units of bits and are indicative of the information content at each position (Schneider and Stephens, 1990; Mount, 2003).
Table II. Residues and length of alignment used for generating profile HMMs

<table>
<thead>
<tr>
<th>Family</th>
<th>Region used for making profile</th>
<th>Total length in alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>ST3Gal</td>
<td>TxxxxYEP-x (20)-FkxDxxW</td>
<td>37</td>
</tr>
<tr>
<td>ST8Sia</td>
<td>KxxxxTnps-x (33)-PAF</td>
<td>46</td>
</tr>
<tr>
<td>ST6Gal</td>
<td>KTTtRxxNSQ-x (22)-WDPxxYx(6)xWxxxxDYxxF</td>
<td>54</td>
</tr>
<tr>
<td>ST6GalNAc subfamily I</td>
<td>SFYGFT-x (78)-HPDFxxYxxxxRF</td>
<td>96</td>
</tr>
<tr>
<td>ST6GalNAc subfamily II</td>
<td>RVxxHxS-x (72)-FxxETGKxR</td>
<td>90</td>
</tr>
</tbody>
</table>

*Residues shown in the sequence logos in Figure 2 are underlined.

The ST3Gal family has 14 sequences, whereas the two subfamilies of ST6GalNAc have six sequences each (Table I). Considering the small size of the data sets, analyses were repeated by considering both experimentally characterized and computationally annotated (details given under Methods) SiaTs. It was found that the motifs are essentially same, but with completely conserved residues being taller in the logos due to the larger data set size (logos are not shown).

The L-motif has been considered to end with the residues VG (Figure S1) in this study. However, this motif has been depicted as having an extension of four, partially conserved, residues in literature (e.g., Harduin-Lepers et al., 2001). This analysis shows that the residues following the L-motif are conserved only at the family level and not at the superfamily level and hence, have been considered separately and not as extensions of the L-motif. Because site-directed mutagenesis studies have implicated the L-motif in donor substrate binding, the newly identified motif that immediately follows the L-motif is more than likely to play a role in conferring linkage specificity. It can also be seen that twoThr residues are present in this motif in both ST3Gal and ST6Gal families. Any possible relationship of this commonality to galactose being the immediate acceptor moiety in these two families needs to be established.

Residues flanking the S-motif are also conserved in the ST8Sia and ST6Gal families (Figure S1). In the ST8Sia family, the S-motif is flanked by an Arg residue and the dipeptide WXF. In the ST6Gal family, the S-motif is flanked by the dipeptide NP and the tetrapeptide PSXR. Such conserved residues are not found in the ST3Gal and ST6GalNAc families.

**ST6GalNAc family comprises two subfamilies**

A closer inspection of the alignment reveals that the ST6GalNAc family consists of two subfamilies, ST6GalNAc III, ST6GalNAc IV, ST6GalNAc V, and ST6GalNAc VI, which resemble each other (pair-wise sequence similarity in the range 49–61%) much more than they resemble either ST6GalNAc I or ST6GalNAc II (pair-wise sequence similarity in the range 18–33%). These four SiaTs thus form one subfamily, whereas ST6GalNAc I and ST6GalNAc II form the other subfamily. These two ST6GalNAc subfamilies have also been recently identified by a different approach, based on phylogeny studies (Harduin-Lepers et al., 2005). Analysis of the sequences by considering only ST6GalNAc I and ST6GalNAc II or only ST6GalNAc III, ST6GalNAc IV, ST6GalNAc V, and ST6GalNAc VI revealed the presence of different linkage-specific sequence motifs, as was found in the ST3Gal, ST8Sia, and ST6Gal families (Figure 2). It has been reported that ST6GalNAc I and ST6GalNAc II utilize sia-lylated as well as unsialylated Galβ-1→3-GalNAc-peptides as acceptors (acceptor saccharide moiety is shown in bold; Kurosawa et al., 1994, 1996), whereas ST6GalNAc III, ST6GalNAc IV, ST6GalNAc V, and ST6GalNAc VI utilize only sialylated acceptor moiety Neu5Ac-α2→3-Galβ-1→3-GalNAc-R (R, glycoprotein or glycolipid) as the acceptor substrate (Lee et al., 1999; Okajima et al., 1999b, 2000b). The relationship of these subfamily-specific motifs with their acceptor substrate preferences needs to be further investigated.

The newly identified linkage-specific motifs are unique to the respective SiaT families

The region of the multiple sequence alignment corresponding to the ST3Gal family-specific L-motif was used as input to derive the profile hidden Markov model (profile HMM; Durbin et al., 1998; Eddy, 1998). This profile was used to query the TrEMBL (1,629,847 sequences, January 2005 release) or the SwissProt (181,821 sequences, May 2005) database. The hits included not only SiaTs of ST3Gal family but also those that belong to other families. A similar result was obtained when the profile HMM for the L-motif of other families or the profile HMM for the S-motif region were used. This indicated that the L- and S-motif region profiles are not specific enough to distinguish the SiaTs belonging to different families. Next, profile HMMs were derived for the regions that lie within the newly identified linkage-specific motifs (Table II). Querying the TrEMBL and SwissProt databases with these profile HMMs resulted in SiaTs belonging to the respective families. This shows that these motifs are indeed linkage- (family-) specific.

**Discussion**

SiaTs from higher organisms comprise a superfamily of enzymes that catalyze the transfer of Sia from the CMP-Sia donor substrate. The presence of the conserved L-, S-, and VS-motifs and motif III in all these SiaTs is indicative of the shared functional features (Gillespie et al., 1992; Wen et al., 1992; Drickamer, 1993; Livingston and Paulson, 1993; Datta and Paulson, 1997; Geremia et al., 1997; Jeanneau et al.,...
The structure–function relationship in SiaTs has been investigated by site-directed mutagenesis in rat (Datta and Paulson, 1995, 1997; Datta et al., 1998, 2001) and human (Laroy et al., 2001) ST6Gal I, human ST3Gal I (Jeanneau et al., 2004), human (Kitazume-Kawaguchi et al., 2001) and murine (Windfuhr et al., 2000) ST8Sia II, and human (Angata et al., 2001; Kitazume-Kawaguchi et al., 2001) and hamster (Windfuhr et al., 2000) ST8Sia IV. Cys4 of L-motif and Cys15 of S-motif (Figure 1) are conserved in all the SiaTs. Mutation of either of these two residues (Cys181 and Cys332 in rat ST6Gal I; Cys142 and C292 in human ST8Sia IV) to Ala results in loss of enzyme activity. The ST8Sia family has a second conserved disulfide bond between C156 (C18 of L-motif; Figure S1) and C356 (near the COOH terminus; after VS motif) and this disulfide bond has also been shown to be essential for the activity of human ST8Sia IV by site-directed mutagenesis studies. However, in the case of other residues, the extent of conservation is not well correlated to the importance of the residue for enzyme activity. For example, mutation of Ser222 to Ala in rat ST6Gal I leads to loss of enzyme activity (<5% of wild type). But this residue is replaced by Thr, Asn, or Gln in other enzymes of the ST6Gal family (residue at position 349 in the multiple alignment in Figure S1). Errors in alignment can be ruled out at this position because the region of the multiple sequence alignment containing Ser222 has no gaps, and the residues that occur before and after Ser222 are well conserved (Figure S1). It is possible that changes at this position are correlated to changes at some other position of the polypeptide chain (i.e., correlated mutations; see, e.g., Fodor and Aldrich, 2004, and references cited therein).

The ST8Sia family appears to have higher sequence conservation, whereas the ST6GalNAc family has the lowest sequence conservation. It has already been recognized that the ST6Gal family is different from the ST6GalNAc family. This analysis shows that even the ST6GalNAc family is heterogeneous and consists of two subfamilies, in agreement with the recent phylogenetic analysis (Harduin-Leper et al., 2005). Each subgroup has characteristic sequence motifs not present in the other subfamily.

Supplementary data
Supplementary data are available at Glycobiology online (http://glycob.oxfordjournals.org/). The multiple sequence alignments of each SiaT family (ST3Gal, ST8Sia, ST6Gal, and ST6GalNAc; Figure S1) are given.

Acknowledgments
The authors thank M.S. Sujatha for helpful discussion and critical reading of the manuscript. R.Y.P. is grateful to Indian Institute of Technology Bombay for teaching assistantship. The authors also thank the anonymous reviewers of the manuscript for their useful comments and suggestions. This work was supported by a grant (No. 37[1110]/02/EMR-II) from the Council for Scientific and Industrial Research, India to P.V.B.
Abbreviations

CAZY, carbohydrate active enzyme; CMP, cytidine monophosphate; HMM, hidden Markov model; L, large; S, small; Sia, sialic acid; SiaTs, sialyltransferases; VS, very small.

References


Linkage-specific motifs in SiaTs

115


