BIOCHEMICAL JOURNAL LETTERS

A classification of nucleotide-diphospho-sugar glycosyltransferases based on amino acid sequence similarities

Glycosyl-transfer reactions are, on quantitative terms, the most important biotransformations on Earth, since they account for the biosynthesis and hydrolysis of the bulk of biomass [1]. The biosynthesis of polysaccharides and complex carbohydrates is also of fundamental biological importance, since these molecules of fascinating diversity directly mediate a wide range of functions, from structure and storage to specific signalling. The biosynthesis of disaccharides, oligosaccharides and polysaccharides involves the action of hundreds of different glycosyltransferases (EC 2.4.*x*.*y*), enzymes which catalyse the transfer of sugar moieties from activated donor molecules to specific acceptor molecules, forming glycosidic bonds. There is a parallel extensive diversity of glycoside hydrolases (EC 3.2.1.*x*), enzymes which cleave such bonds to yield carbohydrates smaller than those from whence they originated. The immense functional and structural variety of glycosyltransferases and glycoside hydrolases raises the problem of their classification.

Regardless of the direction of the reaction, enzymes which catalyse glycosyl-transfer reactions can be classified according to the stereochemistries of the reaction substrates and products as either 'retaining' or 'inverting' enzymes [2]. Furthermore, specific enzymes can be classified on the basis of the reaction catalysed and the substrate specificity, according to the recommendations of the International Union of Biochemistry and Molecular Biology (IUBMB) [3]. However, there are limitations to the utility of this system for classification of glycosyltransferases and glycoside hydrolases, as it does not indicate the intrinsic structural features of the enzymes, nor does it adequately accommodate enzymes which act on several distinct substrates.

Classification of enzymes based on the similarities of their amino acid sequences offers a system complementary to that of the IUBMB [3] and realizes the potential to marry structural features of enzymes with their observed functions. Such classification systems have been proposed for glycoside hydrolases [4] and peptidases [5], and have been updated with the increasing number of cloned genes for these enzymes [6–9]. A significant advantage of classification according to sequence similarities is that it allows logical grouping of enzymes of different EC numbers into polyspecific families and offers insights into the divergent evolution of enzyme families [4]. Conversely, some enzymes which can be grouped by function have been shown to belong to several distinct families and thus reflect convergent evolution [4]. Significantly, the discriminatory power of these classifications has been confirmed by the similarity of the three-dimensional structures [10] and the conserved molecular mechanisms [11] of family members.

Despite the utility of the sequence-based classification of glycoside hydrolases, no such system has been fully described for glycosyltransferases. One difficulty with such a classification is the number of enzymes concerned (193 entries of EC 2.4.1.*x*), and the diversity of sugar donors. These can be di- or polysaccharides, sugar 1-phosphates, or, most commonly, nucleotide diphospho-sugars (NDP-sugars). Whereas a number of the latter type of glycosyltransferases have been compared and grouped into a single family [12], and sequence similarity has been used to predict mechanisms of action [13], there have been no reports of a comprehensive classification of NDP-sugar glycosyltransferases. The present letter describes a classification of NDP-sugar hexosyltransferases (EC 2.4.1.*x*) and related proteins into distinct sequence-based families.

Sequences of NDP-sugar hexosyltransferases were retrieved from the SwissProt and EMBL/GenBank databanks and compiled into a preliminary sequence library which covered the 35 EC 2.4.1.*x* entries for which at least one sequence is known to date. Representatives of each EC number were used as templates for BLAST similarity searches [14], and complementary sequences were retrieved from either SwissProt or EMBL/GenBank. BLAST results were examined using Visual BLAST [15]. When the BLAST probability values were low (typically $P > 10^{-3}$), sequences were further compared by hydrophobic cluster analysis (HCA) [16,17]. A family was defined as a grouping of at least two sequences of significant amino acid or HCA similarity over a length exceeding 100 residues, with no similarity to other families.

A total of 555 sequences were analysed, of which 553 were classified into 26 families (Table 1). Only two sequences, namely those of the mannosyltransferase OCH1 of *Saccharomyces cerevisiae* (GenBank D11095) and the DNA β-glucosyltransferase of bacteriophage T4 (Swiss-Prot P04547), could not be assigned to any family and were left unclassified.

Seven families were found polyspecific (containing two or more EC numbers), whereas the others were either monospecific (one single EC number) or 'uncertain' (no EC numbers assigned to the sequences). Previous experience with the classification of glycoside hydrolases suggests that the number of polyspecific families could increase with the availability of more glycosyltransferase sequences.

More than half of the sequences are found in the three largest families (families 1, 2 and 4 with respectively 107, 139 and 84 members; Table 1). Family 1 comprises proteins from viruses, bacteria, fungi, plants and animals. Families 2, 4, 8 and 20 contain sequences from bacteria, fungi, plants and animals. Conversely, several other small families appear strongly biased toward only one taxonomic group, but this could simply be a consequence of the smaller number of current members in these families.

Sequence similarity is strongly indicative of folding similarity in proteins [18]. Conservation of tertiary structure is such that the same three-dimensional fold is expected to be found within each of the families defined by the present study. For polyspecific families, this suggests that details of the three-dimensional structure, rather than differences in the global fold, will explain different donor and/or acceptor specificities. Whereas to date there has been only one reported three-dimensional structure for a glycosyltransferase, the DNA β-glucosyltransferase of bacteriophage T4 [19], it is inevitable that more of these enzymes will be purified, crystallized and characterized. Family allocation

Table 1 Families of NDP-sugar hexosyltransferases and related protein sequences

The mechanism (retaining or inverting) is indicated for each family where it could be unambiguously identified from the sequence databanks or from the EC recommendations. Notes: [a]Description as found in the sequence databanks; [b]EC number as found in the sequence databanks; [c]accession numbers starting with P or Q are from the Swiss-Prot databank, those starting with PC are from the PIR databank and those starting with other letters are from the EMBL/GenBank databanks. For conciseness, the names of several enzymes have been abbreviated when a known EC number could be given. Specifically, this has been done for EC 2.4.1.101 (β-1,3-mannosyl-glycoprotein β-1,2-*N*-acetylglucosaminyltransferase), EC 2.4.1.102 (β-1,3-galactosyl-*O*-glycosyl-glycoprotein β-1,6-*N*-acetylglucosaminyltransferase), EC 2.4.1.143 (α-1,6-mannosyl-glycoprotein β-1,2-*N*-acetylglucosaminyltransferase), EC 2.4.1.144 (β-1,4-mannosyl-glycoprotein β-1,4-*N*-acetylglucosaminyl-transferase), and EC 2.4.1.155 [α-1,3(6)-mannosyl-glycoprotein β-1,6-*N*-acetylglucosaminyltransferase]. Abbreviations used: GPI, glycosyl-phosphatidylinositol; LPS, lipopolysaccharide; GlcNAc, *N*-acetylglucosamine.

Description[a]	EC number[b]	Organism	Accession no.[c]
Family 1 (inverting)			
Glycosyltransferase GtfA		*Amycolatopsis orientalis*	U84349
Glycosyltransferase GtfB		*Amycolatopsis orientalis*	U84349
Glycosyltransferase GtfC		*Amycolatopsis orientalis*	U84349
Glycosyltransferase GtfD		*Amycolatopsis orientalis*	U84350
Glycosyltransferase GtfE		*Amycolatopsis orientalis*	U84350
Ecdysteroid glucosyltransferase		*Autographa californica* NP virus	P18569
Unknown AC3.2		*Caenorhabditis elegans*	Z71177
Unknown AC3.7		*Caenorhabditis elegans*	Z71177
Unknown AC3.8		*Caenorhabditis elegans*	Z71177
Unknown B0310.5		*Caenorhabditis elegans*	U40959
Unknown C07A9.6		*Caenorhabditis elegans*	P34317
Unknown C08B6.1		*Caenorhabditis elegans*	Z72502
Unknown C17G1.3		*Caenorhabditis elegans*	Z78415
Unknown C18C4.3		*Caenorhabditis elegans*	U55369
Unknown C23G10.6		*Caenorhabditis elegans*	U39851
Unknown C33A12.6		*Caenorhabditis elegans*	Z68493
Unknown C35A5.2		*Caenorhabditis elegans*	Z71185
Unknown C44H9.1		*Caenorhabditis elegans*	Z75529
Unknown C55H1.1		*Caenorhabditis elegans*	U55367
Unknown F01E11.1		*Caenorhabditis elegans*	U42832
Unknown F08G5.5		*Caenorhabditis elegans*	Z70682
Unknown F29F11.2		*Caenorhabditis elegans*	Z73905
Unknown F35H8.6		*Caenorhabditis elegans*	Z36752
Unknown R04B5.9		*Caenorhabditis elegans*	Z70782
Unknown R11A8.3		*Caenorhabditis elegans*	Z70310
Unknown T04H1.7		*Caenorhabditis elegans*	Z78200
Unknown T04H1.8		*Caenorhabditis elegans*	Z78200
Unknown T07C5.1		*Caenorhabditis elegans*	Z50006
Unknown T25B9.7		*Caenorhabditis elegans*	Z70311
Unknown ZC443.6		*Caenorhabditis elegans*	Z75553
Unknown ZC455.3		*Caenorhabditis elegans*	Z75554
Unknown ZC455.4		*Caenorhabditis elegans*	Z75554
Unknown ZC455.5		*Caenorhabditis elegans*	Z75554
Unknown ZC455.6		*Caenorhabditis elegans*	Z75554
Ecdysteroid glucosyltransferase		*Choristoneura fumiferana* NP virus 1	U10441
Ecdysteroid glucosyltransferase		*Choristoneura fumiferana* NP virus 2	U10476
CrtX protein		*Erwinia herbicola*	M90698
Zeaxanthin glucosyltransferase		*Erwinia herbicola*	Q01330
Zeaxanthin glucosyltransferase		*Erwinia uredovora*	P21686
Flavonol O^3-glucosyltransferase	2.4.1.91	*Gentiana triflora*	D85186
Flavonol O^3-glucosyltransferase	2.4.1.91	*Hordeum vulgare*	P14726
1-β-Galactosyltransferase	2.4.1.45	Human	U32370
Glucuronosyltransferase 1B	2.4.1.17	Human	P36509
Glucuronosyltransferase 1C	2.4.1.17	Human	P35503
Glucuronosyltransferase 1D	2.4.1.17	Human	P22310
Glucuronosyltransferase 1E	2.4.1.17	Human	P35504
Glucuronosyltransferase 1A	2.4.1.17	Human	P22309
Glucuronosyltransferase 1F	2.4.1.17	Human	P19224
Glucuronosyltransferase 2B10	2.4.1.17	Human	P36537
Glucuronosyltransferase 2B11	2.4.1.17	Human	P36538
Glucuronosyltransferase 2B4	2.4.1.17	Human	P06133
Glucuronosyltransferase 2B7	2.4.1.17	Human	P16662
Glucuronosyltransferase 2B8	2.4.1.17	Human	P23765
Glucuronosyltransferase 2B15	2.4.1.17	Human	P54855
Ecdysteroid glucosyltransferase		*Lacanobia oleracea* granulosis virus	Y08294
Twi1 protein		*Lycopersicon esculentum*	X85138
Ecdysteroid glucosyltransferase		*Lymantria dispar* multicapsid NP virus	P41713
Ecdysteroid glucosyltransferase		*Mamestra brassicae* NP virus	U41999
UDP-glucose glucosyltransferase Cgt1		*Manihot esculenta*	X77459
UDP-glucose glucosyltransferase Cgt5		*Manihot esculenta*	X77462
Galactosyltransferase	2.4.1.62	Mouse	X92122
Glucuronosyltransferase 1A1	2.4.1.17	Mouse	D87867

Table 1 (cont.)

Description[a]	EC number[b]	Organism	Accession no.[c]
Glucuronosyltransferase 1-06	2.4.1.17	Mouse	U16818
Glucuronosyltransferase UGTP4	2.4.1.17	Mouse	L27122
Glucuronosyltransferase UGTBr1	2.4.1.17	Mouse	S64760
Glucuronosyltransferase 2B5	2.4.1.17	Mouse	P17717
Unknown (PID:E256387)		*Mycobacterium tuberculosis*	Z77826
Unknown (PID:E256535)		*Mycobacterium tuberculosis*	Z77826
Salicylate-induced glucosyltransferase IS10a		*Nicotiana tabacum*	U32643
Salicylate-induced glucosyltransferase IS5a		*Nicotiana tabacum*	U32644
UDP-rhamnose rhamnosyltransferase		*Petunia hybrida*	Z25802
Glucuronosyltransferase		*Pleuronectes platessa*	X74116
Rhamnosyl transferase		*Pseudomonas aeruginosa*	L28170
Glucuronosyltransferase UGT1-6		Rabbit	U09030
Glucuronosyltransferase UGT1-4		Rabbit	U09101
Glucuronosyltransferase 2A2		Rabbit	P36514
Glucuronosyltransferase 2B13		Rabbit	P36512
Glucuronosyltransferase 2B14		Rabbit	P36513
1-β-Galactosyltransferase	2.4.1.45	Rat	Q09426
Glucuronosyltransferase 1.1	2.4.1.17	Rat	U20551
Glucuronosyltransferase		Rat	J05132
Glucuronosyltransferase 2B12		Rat	P36511
Glucuronosyltransferase		Rat	U27518
Glucuronosyltransferase UGT1		Rat	D38063
Glucuronosyltransferase 1A		Rat	P20720
Glucuronosyltransferase 1F		Rat	P08430
Glucuronosyltransferase 2A1		Rat	P36510
Glucuronosyltransferase 2B1		Rat	P09875
Glucuronosyltransferase 2B2		Rat	P08541
Glucuronosyltransferase 2B3		Rat	P08542
Glucuronosyltransferase 2B6		Rat	P19488
Glucuronosyltransferase R-21		Rat	P19489
Unknown (gene L9470.23)		*Saccharomyces cerevisiae*	U17246
Glycosyltransferase		*Solanum melongena*	X77369
Glucuronosyltransferase		*Solanum tuberosum*	U82367
Ecdysteroid glucosyltransferase		*Spodoptera littoralis* NP virus	X84701
Glycosyltransferase		*Streptomyces antibioticus*	Z22577
Glycosyltransferase		*Streptomyces fradiae*	X81885
Glycosyltransferase		*Streptomyces peucetius*	L47164
Macrolide glycosyltransferase		*Streptomyces lividans*	M74717
Glycosyltransferase		*Streptomyces* sp. C5	U43704
Zeaxanthin glucosyltransferase CrtX		*Synechocystis* sp.	D90899
Flavonol O^3-glucosyltransferase	2.4.1.91	*Vitis vinifera*	P51094
Flavonol O^3-glucosyltransferase 1	2.4.1.91	*Zea mays*	P16165
Flavonol O^3-glucosyltransferase 2	2.4.1.91	*Zea mays*	P16166
Flavonol O^3-glucosyltransferase 3	2.4.1.91	*Zea mays*	P16167
Indole-3-acetate β-glucosyltransferase	2.4.1.121	*Zea mays*	L34847
Family 2 (inverting)			
Cellulose synthase AcsA	2.4.1.12	*Acetobacter xylinum*	P19449
Cellulose synthase AcsAII	2.4.1.12	*Acetobacter xylinum*	U15957
Cellulose synthase BcsA	2.4.1.12	*Acetobacter xylinum*	P21877
Cellulose synthase	2.4.1.12	*Agrobacterium tumefaciens*	L38609
Unknown		*Anabaena* sp.	P22639
Chitin synthase C	2.4.1.16	*Aspergillus fumigatus*	X94245
Chitin synthase G-1	2.4.1.16	*Aspergillus fumigatus*	U39478
Chitin synthase G-2	2.4.1.16	*Aspergillus fumigatus*	U39479
GlcNAc transferase (NodC)		*Azorhizobium caulinodans*	Q07755
CgeBB protein		*Bacillus subtilis*	P42092
CsbB protein		*Bacillus subtilis*	L77099
Teichoic acid biosynthesis protein GgaA		*Bacillus subtilis*	P46917
Teichoic acid biosynthesis protein GgaB		*Bacillus subtilis*	P46918
SpsA protein		*Bacillus subtilis*	P39621
YwdF protein		*Bacillus subtilis*	P39614
YveO protein		*Bacillus subtilis*	Z71928
Polypeptide GalNAc transferase	2.4.1.41	Bovine	Q07537
GlcNAc transferase (NodC)		*Bradyrhizobium elkanii*	D28963
GlcNAc transferase (NodC)		*Bradyrhizobium* sp.	P04677
Unknown F13G3.6		*Caenorhabditis elegans*	Z71259
Unknown ZK688.8		*Caenorhabditis elegans*	P34678
Chitin synthase	2.4.1.16	*Candida albicans*	P30572
Chitin synthase	2.4.1.16	*Candida albicans*	P30573
Chitin synthase	2.4.1.16	*Candida maltosa*	D29762

Table 1 (cont.)

Description[a]	EC number[b]	Organism	Accession no.[c]
Chitin oligosaccharide synthase		*Danio rerio*	U53223
Chitin synthase A	2.4.1.16	*Emericella nidulans*	D21268
Chitin synthase B	2.4.1.16	*Emericella nidulans*	D21269
Chitin synthase D	2.4.1.16	*Emericella nidulans*	U62895
Chitin synthase E	2.4.1.16	*Emericella nidulans*	U52362
AmsB protein		*Erwinia amylovora*	X77921
AmsE protein		*Erwinia amylovora*	X77921
KfiC protein		*Escherichia coli*	X77617
WcaA protein		*Escherichia coli*	U38473
WcaE protein		*Escherichia coli*	U38473
Dolichol-phosphate mannosyltransferase	2.4.1.83	*Escherichia coli*	D90856
YhjO protein		*Escherichia coli*	P37653
Unknown (ORF275)		*Escherichia coli*	L04596
Unknown protein F344		*Escherichia coli*	P11290
Cellulose synthase	2.4.1.12	*Gossypium hirsutum*	U58283
Unknown HI0868		*Haemophilus influenzae*	U32768
Unknown HI1578		*Haemophilus influenzae*	U32832
Unknown HI1695		*Haemophilus influenzae*	U32842
Unknown HI1696		*Haemophilus influenzae*	U32842
Unknown (ORF5)		*Haemophilus influenzae*	M94855
Unknown (ORF6)		*Haemophilus influenzae*	M94855
Hyaluronan synthase 1		Human	D84424
Hyaluronan synthase 2		Human	U59269
Polypeptide GalNAc transferase T1	2.4.1.41	Human	Q10472
Polypeptide GalNAc transferase T2	2.4.1.41	Human	X85019
Polypeptide GalNAc transferase T3	2.4.1.41	Human	X92689
EpsG protein		*Lactococcus lactis cremoris*	U93364
Dolichol-phosphate mannosyltransferase	2.4.1.83	*Methanococcus jannaschii*	U67504
Unknown (MJ1057)		*Methanococcus jannaschii*	U67549
Chitin oligosaccharide synthase		Mouse	U53222
Hyaluronan synthase 1		Mouse	D82964
Hyaluronan synthase 2		Mouse	U52524
Polypeptide GalNAc transferase	2.4.1.41	Mouse	U70538
Unknown (PID:G699128)		*Mycobacterium leprae*	U15180
Unknown (PID:E264129)		*Mycobacterium tuberculosis*	Z79701
Unknown (PID:E264145)		*Mycobacterium tuberculosis*	Z79701
Unknown (PID:E264147)		*Mycobacterium tuberculosis*	Z79701
Unknown (PID:E256390)		*Mycobacterium tuberculosis*	Z77826
Unknown (PID:E283381)		*Mycobacterium tuberculosis*	Z83018
RfbV protein		*Mycoplasma genitalium*	U39685
TrsB protein		*Mycoplasma genitalium*	P47271
RfbC protein		*Myxococcus xanthus*	U36795
GlcNAc transferase LgtA		*Neisseria gonorrhoeae*	U14554
GalNAc transferase LgtD		*Neisseria gonorrhoeae*	U14554
Glycosyltransferase LgtA		*Neisseria meningitidis*	U25839
Chitin synthase 1	2.4.1.16	*Neurospora crassa*	P30588
Chitin synthase 2	2.4.1.16	*Neurospora crassa*	P30589
Chitin synthase 3	2.4.1.16	*Neurospora crassa*	P29070
Chitin synthase 4	2.4.1.16	*Neurospora crassa*	Q01285
Unknown		*Paramecium bursaria Chlorella* virus 1	U42580
Polypeptide GalNAc transferase	2.4.1.41	Pig	D85389
Unknown (ORF1)		*Porphyromonas gingivalis*	U60208
Alginate-synthesis related protein Alg8		*Pseudomonas aeruginosa*	L22611
MigA protein		*Pseudomonas aeruginosa*	U70729
Polypeptide GalNAc transferase	2.4.1.41	Rat	Q10473
GlcNAc transferase (NodC)		*Rhizobium galegae*	P50356
GlcNAc transferase (NodC)		*Rhizobium leguminosarum* (*viciae*)	P04340
GlcNAc transferase (NodC)		*Rhizobium leguminosarum* (*phaseoli*)	P24151
GlcNAc transferase (NodC)		*Rhizobium loti*	P17862
ExoM protein		*Rhizobium meliloti*	P33695
ExoO protein		*Rhizobium meliloti*	P33697
ExoU protein		*Rhizobium meliloti*	P33700
ExoW protein		*Rhizobium meliloti*	P33702
GlcNAc transferase (NodC)		*Rhizobium meliloti*	P04341
GlcNAc transferase (NodC)		*Rhizobium* sp. strain N33	U53327
GlcNAc transferase (NodC)		*Rhizobium* sp. strain NGR234	P50357
GlcNAc transferase (NodC)		*Rhizobium tropici*	X98514
Chitin synthase	2.4.1.16	*Rhizopus oligosporus*	P30594
Unknown		*Rhodococcus* sp.	P46370
Chitin synthase 1	2.4.1.16	*Saccharomyces cerevisiae*	P08004
Chitin synthase 2	2.4.1.16	*Saccharomyces cerevisiae*	P14180

Table 1 (cont.)

Description[a]	EC number[b]	Organism	Accession no.[c]
Chitin synthase 3	2.4.1.16	*Saccharomyces cerevisiae*	P29465
Dolichol-phosphate mannosyltransferase	2.4.1.83	*Saccharomyces cerevisiae*	P14020
Dolichyl-phosphate β-glucosyltransferase	2.4.1.117	*Saccharomyces cerevisiae*	P40350
Rhamnosyltransferase RfbQ		*Salmonella enterica*	X61917
Rhamnosyltransferase WbaN		*Salmonella enterica*	X60665
N-Acetylmannosamine transferase RfbA		*Salmonella enterica* (plasmid pWQ799)	L39794
Rhamnosyltransferase RfbN		*Salmonella typhimurium*	P26403
RfbV protein		*Salmonella typhimurium*	P26401
Unknown (ORF 14.1)		*Salmonella typhimurium*	M65054
Chitin synthase	2.4.1.16	*Saprolegnia monoica*	U19946
Chitin synthase (class VI)	2.4.1.16	*Sartorya fumigata*	U62614
KdtX protein		*Serratia marcescens*	U52844
RfpA protein		*Shigella dysenteriae*	S73325
RfbG protein		*Shigella flexneri*	X71970
Unknown (ORF7)		*Shigella sonnei*	U34305
SpsL protein		*Sphingomonas* S88	U51197
IcaA protein		*Staphylococcus epidermidis*	U43366
Fbf15 protein		*Stigmatella aurantiaca*	Z11601
Cellobiouronic acid synthase		*Streptococcus pneumoniae*	Z47210
Type 3 capsular polysaccharide synthase		*Streptococcus pneumoniae*	U15171
Hyaluronan synthase 1		*Streptococcus pyogenes*	L20853
Hyaluronan synthase		*Streptococcus pyogenes* strain WF14	L21187
EpsI protein		*Streptococcus thermophilus*	U40830
Unknown (ORF C04008)		*Sulfolobus solfataricus*	Y08257
Unknown		*Synechococcus* sp.	P42460
Unknown (ORF SLL1664)		*Synechocystis* sp.	D90900
Unknown (ORF SLL1020)		*Synechocystis* sp.	D90901
Unknown (ORF SLR1537)		*Synechocystis* sp.	D90906
SpsA protein		*Synechocystis* sp.	D90911
Unknown (ORF SLR2120)		*Synechocystis* sp.	D90911
Unknown (ORF SLL1377)		*Synechocystis* sp.	D90912
Dolichol-phosphate mannosyltransferase	2.4.1.83	*Trypanosoma brucei*	Z54162
Dolichol-phosphate mannosyltransferase	2.4.1.83	*Ustilago maydis*	P54856
VirB protein		*Vibrio anguillarum*	L08012
Glycosyltransferase B (putative)		*Vibrio cholerae*	U72485
DG42 protein		*Xenopus laevis*	P13563
RfbB protein		*Yersinia enterocolitica*	Z18920
RfbH protein		*Yersinia enterocolitica*	Z18920
TrsB protein		*Yersinia enterocolitica*	Z47767
TrsC protein		*Yersinia enterocolitica*	Z47767
Unknown (ORF7.8)		*Yersinia enterocolitica*	U46859
Unknown (ORF10.9)		*Yersinia enterocolitica*	U46859
HmsR protein		*Yersinia pestis*	U22837
Abequosyltransferase (putative)		*Yersinia pseudotuberculosis*	L01777
Family 3 (retaining)			
Glycogen synthase	2.4.1.11	Human	P13807
Glycogen synthase	2.4.1.11	Mouse	P54859
Glycogen synthase	2.4.1.11	Rabbit	P13834
Glycogen synthase	2.4.1.11	Rat	P17625
Glycogen synthase isoform 1	2.4.1.11	*Saccharomyces cerevisiae*	P23337
Glycogen synthase isoform 2	2.4.1.11	*Saccharomyces cerevisiae*	P27472
Family 4 (retaining)			
AceC protein		*Acetobacter xylinum*	X94981
Sucrose synthase	2.4.1.13	*Alnus glutinosa*	P49034
HepB protein		*Anabaena* sp.	U68035
Sucrose synthase	2.4.1.13	*Arabidopsis thaliana*	P49040
Techoic acid biosynthesis protein E	2.4.1.52	*Bacillus subtilis*	P13484
Unknown		*Bacillus subtilis*	P46915
Unknown		*Bacillus subtilis*	P42982
YveP protein		*Bacillus subtilis*	Z71928
YqgM protein		*Bacillus subtilis*	D84432
Sucrose-phosphate synthase	2.4.1.14	*Beta vulgaris*	P49031
Sucrose synthase SBSS1	2.4.1.13	*Beta vulgaris*	X81974
BplH protein		*Bordetella pertussis*	X90711
Galactosyltransferase		*Campylobacter hyoilei*	X91081
AmsD protein		*Erwinia amylovora*	X77921
AmsK protein		*Erwinia amylovora*	X77921
LPS 1,6-galactosyltransferase RfaB		*Escherichia coli*	P27127
LPS core biosynthesis protein RfaG		*Escherichia coli*	P25740

Table 1 (cont.)

Description[a]	EC number[b]	Organism	Accession no.[c]
Mannosyltransferase A		*Escherichia coli*	D43637
Mannosyltransferase B		*Escherichia coli*	D43637
Mannosyltransferase C		*Escherichia coli*	D43637
O-antigen biosynthesis protein rfb		*Escherichia coli*	X59852
WcaC protein		*Escherichia coli*	U38473
WcaL protein		*Escherichia coli*	U38473
Unknown (ORF3)		*Haemophilus influenzae*	U36398
Sucrose synthase 1	2.4.1.13	*Hordeum vulgare*	P31922
Sucrose synthase 2	2.4.1.13	*Hordeum vulgare*	P31923
GlcNAc-phosphatidylinositol transferase		Human	P37287
Galactosyltransferase RfbF		*Klebsiella pneumoniae*	L41518
Unknown (ORF7)		*Klebsiella pneumoniae*	D21242
Unknown (ORF8)		*Klebsiella pneumoniae*	D21242
Capsular LPS biosynthesis protein M		*Methanococcus jannaschii*	U67549
Unknown (MJ1185)		*Methanococcus jannaschii*	U67559
LPS biosynthesis related RfbU-protein		*Methanococcus jannaschii*	U67601
GPI anchor biosynthesis protein PigA		Mouse	D26047
Unknown (U2168F)		*Mycobacterium leprae*	P54138
Unknown (MTCY20G9.12)		*Mycobacterium tuberculosis*	Q11152
Unknown (MTCY25D10.36)		*Mycobacterium tuberculosis*	Z95558
α-1,2-GlcNAc transferase RfaK		*Neisseria meningitidis*	U35713
LPS yyscLPS glycosyltransferase IcsA		*Neisseria meningitidis*	U39810
Sucrose synthase 1	2.4.1.13	*Oryza sativa*	P30298
Sucrose synthase 3	2.4.1.13	*Oryza sativa*	L03366
Sucrose-phosphate synthase	2.4.1.14	*Phaseolus aureus*	Q01390
Sucrose synthase	2.4.1.13	*Pisum sativum*	X98598
CpsF protein		*Proteus mirabilis*	L36873
GlcNAc-phosphatidylinositol transferase		*Saccharomyces cerevisiae*	P32363
Unknown (ORF YPL175w)		*Saccharomyces cerevisiae*	Z73531
Mannosyltransferase WbaW		*Salmonella enterica*	X61917
Mannosyltransferase WbaZ		*Salmonella enterica*	X61917
LPS GlcNAc transferase RfaK	2.4.1.56	*Salmonella typhimurium*	P26470
Mannosyltransferase RfbU		*Salmonella typhimurium*	P26402
Polysaccharide biosynthesis protein VipC		*Salmonella typhimurium*	Q04975
Galactosyltransferase RfbF		*Serratia marcescens*	L34167
Galactosyltransferase RfpB		*Shigella dysenteriae*	S73325
ExpA3 protein		*Sinorhizobium meliloti*	Z79692
Sucrose-phosphate synthase	2.4.1.14	*Solanum tuberosum*	X73477
Sucrose synthase	2.4.1.13	*Solanum tuberosum*	P10691
Sucrose-phosphate synthase	2.4.1.14	*Spinacia oleracea*	P31928
CapM protein		*Staphylococcus aureus*	P39862
Cap1G protein		*Streptococcus pneumoniae*	Z83335
EpsF protein		*Streptococcus thermophilus*	U40830
EpsG protein		*Streptococcus thermophilus*	U40830
LPS glycosyltransferase IcsA		*Synechocystis* sp.	D90906
Mannosyltransferase RfbW		*Synechocystis* sp.	D64000
Mannosyltransferase RfbU		*Synechocystis* sp.	D90901
Mannosyltransferase MtfB		*Synechocystis* sp.	D90911
Unknown (ORF SLR1508)		*Synechocystis* sp.	D90911
Unknown (ORF SLR0384)		*Synechocystis* sp.	D63999
Sucrose-phosphate synthase (ORF SLL0045)		*Synechocystis* sp.	D64006
Unknown (ORF SLR1077)		*Synechocystis* sp.	D90901
Unknown (ORF SLR1065)		*Synechocystis* sp.	D90901
Unknown (ORF SLL1971)		*Synechocystis* sp.	D90905
Unknown (ORF SLL1723)		*Synechocystis* sp.	D90906
Unknown (ORF SLR1166)		*Synechocystis* sp.	D90913
Sucrose synthase type 1	2.4.1.13	*Triticum aestivum*	M26671
Sucrose synthase type 2	2.4.1.13	*Triticum aestivum*	M26672
Sucrose synthase	2.4.1.13	*Tulipa gesneriana*	X96939
Sucrose-phosphate synthase	2.4.1.14	*Vicia faba*	Z56278
GumH protein		*Xanthomonas campestris*	U22511
RfbC protein		*Yersinia enterocolitica*	Z18920
RfbP protein		*Yersinia enterocolitica*	U46859
TrsD protein		*Yersinia enterocolitica*	Z47767
TrsE protein		*Yersinia enterocolitica*	Z47767
TrsH protein		*Yersinia enterocolitica*	Z47767
Sucrose synthase 1	2.4.1.13	*Zea mays*	P04712
Sucrose synthase 2	2.4.1.13	*Zea mays*	P49036
Family 5 (retaining)			
Bacterial glycogen synthase	2.4.1.21	*Agrobacterium tumefaciens*	P39670

Table 1 (cont.)

Description[a]	EC number[b]	Organism	Accession no.[c]
Bacterial glycogen synthase	2.4.1.21	*Bacillus subtilis*	P39125
Bacterial glycogen synthase	2.4.1.21	*Escherichia coli*	P08323
Bacterial glycogen synthase (putative)		*Haemophilus influenzae*	P45179
Starch synthase	2.4.1.21	*Hordeum vulgare*	P09842
Starch synthase	2.4.1.21	*Ipomoea batatas*	U44126
Starch synthase	2.4.1.21	*Manihot esculenta*	X74160
Bacterial glycogen synthase	2.4.1.21	*Methanococcus jannaschii*	U67600
Glucosyl transferase		*Oryza glaberrima*	D10472
Starch synthase (soluble)	2.4.1.21	*Oryza sativa*	D16202
Starch synthase	2.4.1.21	*Oryza sativa*	P19395
Starch synthase (granule-bound)	2.4.1.21	*Oryza sativa*	X64108
Starch synthase (WaxyG)	2.4.1.21	*Oryza sativa*	X65183
Starch synthase (isoform 1)	2.4.1.21	*Pisum sativum*	X88790
Starch synthase (isoform 2)	2.4.1.21	*Pisum sativum*	X88789
Starch synthase (isoform 1)	2.4.1.11	*Solanum tuberosum*	Q00775
Starch synthase (isoform 2)	2.4.1.21	*Solanum tuberosum*	X83220
Starch synthase (isoform 3)	2.4.1.21	*Solanum tuberosum*	X95759
Starch synthase (isoform 4)	2.4.1.21	*Solanum tuberosum*	X87988
Starch synthase	2.4.1.21	*Sorghum bicolor*	U23945
Unknown (ORF SLL1393)		*Synechocystis* sp.	D90899
Unknown (ORF SLL0945)		*Synechocystis* sp.	D90915
Starch synthase (isoform 1)	2.4.1.11	*Triticum aestivum*	P27736
Starch synthase (soluble)	2.4.1.21	*Triticum aestivum*	U48227
Starch synthase (isoform 2)	2.4.1.21	*Triticum aestivum*	U66377
Starch synthase	2.4.1.21	*Zea mays*	P04713
Family 6 (retaining)			
Histo-blood-group-A transferase		Baboon	PC1172
Histo-blood-group-B transferase		Baboon	PC1173
3-α-Galactosyltransferase	2.4.1.124	Bovine	P14769
Histo-blood-group-2 transferase		Chimpanzee	PC1166
Histo-blood-group transferase		Crab-eating macaque	PC1171
Histo-blood-group transferase		Dog	U66140
Histo-blood-group-2 transferase		Gorilla	PC1168
Blood-group-B α-1,3-galactosyltransferase		Human	X91874
Fucosylglycoprotein α-GalNAc transferase		Human	P16442
Histo-blood-group-A transferase		Human	X84746
Histo-blood-group-A transferase		Human	J05175
Histo-blood-group-B transferase		Human	PC1165
α-1,3-Galactosyltransferase	2.4.1.151	Marmoset	S71333
α-1,3-Galactosyltransferase		Mouse	M85153
α-1,3-Galactosyltransferase	2.4.1.151	Mouse	P23336
Histo-blood-group-1 transferase		Orang-utan	PC1169
Histo-blood-group-2 transferase		Orang-utan	PC1170
α-1,3-Galactosyltransferase 1		Pig	L36535
α-1,3-Galactosyltransferase 2		Pig	P50127
Family 7 (inverting)			
Glycoprotein 4-β-galactosyltransferase	2.4.1.38/90	Bovine	P08037
N-Acetyl-lactosamine synthase		*Caenorhabditis elegans*	X98132
Unknown R10E11.4		*Caenorhabditis elegans*	Z29095
Unknown W02B12.11		*Caenorhabditis elegans*	Z66521
β-1,4-Galactosyltransferase CKI		Chicken	U19890
β-1,4-Galactosyltransferase CKII		Chicken	U19889
β-1,4-Galactosyltransferase		Human	D29805
β-1,4-Galactosyltransferase		Human	U10473
β-1,4-Galactosyltransferase Gtn2		Human	U10472
β-1,4-Galactosyltransferase Gtn6		Human	U10474
Glycoprotein 4-β-galactosyltransferase	2.4.1.38/90	Human	P15291
β-1,4-Galactosyltransferase (isoform 1)		*Lymnaea stagnalis*	X99318
β-1,4-Galactosyltransferase (isoform 2)		*Lymnaea stagnalis*	X99319
β-1,4-GlcNAc transferase		*Lymnaea stagnalis*	X80228
Glycoprotein 4-β-galactosyltransferase	2.4.1.38	Mouse	P15535
β-1,4-Galactosyltransferase		Pig	U63019
Family 8 (retaining)			
Unknown (ORF2)		*Bacillus subtilis*	P25148
Unknown F56B6.4		*Caenorhabditis elegans*	U64599
Unknown T10B10.8		*Caenorhabditis elegans*	Z72514
LPS α-1,2-galactosyltransferase	2.4.1.58	*Escherichia coli*	P27129
LPS α-1,3-galactosyltransferase	2.4.1.44	*Escherichia coli*	P27128

Table 1 (cont.)

Description[a]	EC number[b]	Organism	Accession no.[c]
α-Galactosyltransferase LgtC		*Haemophilus influenzae*	P44597
Glycogenin	2.4.1.186	Human	P46976
RfbC protein 1		*Klebsiella pneumoniae*	L31762
RfbC protein 2		*Klebsiella pneumoniae*	L41518
Glycosyltransferase LgtC		*Neisseria gonorrhoeae*	U14554
Glycosyltransferase LgtC		*Neisseria meningitidis*	U65788
WSI76 water-stress protein		*Oryza sativa*	D26537
Glycogenin	2.4.1.186	Rabbit	P13280
Galactosyltransferase LgtA		*Rhizobium leguminosarum*	X94963
Unknown (PID:G152040)		*Rhodobacter sphaeroides*	M89780
Glycogenin		*Saccharomyces cerevisiae*	P36143
Unknown 44.5 kDa protein		*Saccharomyces cerevisiae*	P47011
Unknown YKRO58W		*Saccharomyces cerevisiae*	Z28283
LPS α-1,2-galactosyltransferase	2.4.1.58	*Salmonella typhimurium*	P19817
LPS α-1,3-galactosyltransferase	2.4.1.44	*Salmonella typhimurium*	P19816
Family 9 (inverting)			
RfaC protein		*Bordetella pertussis*	X90711
LPS 1,2-GlcNAc transferase	2.4.1.56	*Escherichia coli*	P27242
LPS heptosyltransferase RfaF		*Escherichia coli*	P37692
LPS heptosyltransferase RfaC		*Escherichia coli*	P24173
LPS biosynthesis protein LbgB		*Haemophilus ducreyi*	U58147
OpsX protein		*Haemophilus influenzae*	U32712
RfaF protein		*Haemophilus influenzae*	L76100
Unknown HI0523		*Haemophilus influenzae*	P44011
RfaC protein		*Neisseria gonorrhoeae*	U10385
RfaC protein 1		*Neisseria meningitidis*	U40862
RfaC protein 2		*Neisseria meningitidis*	U35454
RfaC protein		*Salmonella typhimurium*	P26469
Family 10 (inverting)			
α-(1,3/4)-Fucosyltransferase 3		Bovine	Q11126
Unknown T05A7.5 (tandem repeat)		*Caenorhabditis elegans*	U40028
Unknown K08F8.3		*Caenorhabditis elegans*	Z66497
α-1,3-Fucosyltransferase		Chicken	U73678
Galactoside 3(4)-L-fucosyltransferase 3	2.4.1.65	Human	P21217
α-1,3-Fucosyltransferase 4		Human	P22083
α-1,3-Fucosyltransferase 5		Human	Q11128
α-1,3-Fucosyltransferase 6		Human	P51993
α-1,3-Fucosyltransferase 7		Human	Q11130
α-1,3/4-Fucosyltransferase		Human	D89325
α-1,3-Fucosyltransferase 4		Mouse	Q11127
α-1,3-Fucosyltransferase 7		Mouse	Q11131
α-1,3-Fucosyltransferase		Rat	U58860
Family 11 (inverting)			
FUT2 gene product		Bovine	X99620
Unknown C06E1.7		*Caenorhabditis elegans*	L16559
Unknown F17B5.e		*Caenorhabditis elegans*	Z81066
Galactoside 2-α-L-fucosyltransferase 1	2.4.1.69	Human	P19526
α-1,2-Fucosyltransferase 2		Human	D82933
α-1,2-Fucosyltransferase Se2		Human	D89327
α-1,2-Fucosyltransferase Sec2		Human	Q10981
Galactoside 2-α-L-fucosyltransferase	2.4.1.69	Mouse	Y09882
α-1,2-Fucosyltransferase		Pig	L50534
FUT2 gene product		Pig	X99621
Galactoside 2-α-L-fucosyltransferase 1	2.4.1.69	Rabbit	Q10979
Galactoside 2-α-L-fucosyltransferase 2	2.4.1.69	Rabbit	Q10983
Galactoside 2-α-L-fucosyltransferase 3	2.4.1.69	Rabbit	X91269
Galactoside 2-α-L-fucosyltransferase 1	2.4.1.69	Rat	Q10980
Galactoside 2-α-L-fucosyltransferase 2	2.4.1.69	Rat	Q10984
Unknown (ORF11.8)		*Yersinia enterocolitica*	U46859
Family 12 (inverting)			
β-1,4-GalNAc transferase	2.4.1.92	Human	Q00973
β-1,4-GalNAc transferase 1	2.4.1.92	Mouse	Q09199
β-1,4-GalNAc transferase 2	2.4.1.92	Mouse	Q09200
β-1,4-GalNAc transferase	2.4.1.92	Rat	Q10468
Family 13 (inverting)			
Unknown B0416.6		*Caenorhabditis elegans*	U23516
Unknown F48E3.1		*Caenorhabditis elegans*	U28735

Table 1 (cont.)

Description[a]	EC number[b]	Organism	Accession no.[c]
Unknown M01F1.1		*Caenorhabditis elegans*	Z46381
GlcNAc transferase I		*Cricetulus griseus*	U65791
β-1,2-GlcNAc transferase I	2.4.1.101	Human	P26572
β-1,2-GlcNAc transferase I	2.4.1.101	Mouse	P27808
β-1,2-GlcNAc transferase I	2.4.1.101	Rabbit	P27115
β-1,2-GlcNAc transferase I	2.4.1.101	Rat	Q09325
Family 14 (inverting)			
β-1,6-GlcNAc transferase	2.4.1.102	Bovine	U41320
Unknown T14B4.9		*Caenorhabditis elegans*	U50191
Unknown F22D6.11		*Caenorhabditis elegans*	Z71262
Unknown F30A10.4		*Caenorhabditis elegans*	Z81072
Unknown F44F4.6		*Caenorhabditis elegans*	Z37092
Unknown R07B7.6		*Caenorhabditis elegans*	Z75955
Unknown T09E11.a		*Caenorhabditis elegans*	Z81147
β-1,6-GlcNAc transferase C2GNT	2.4.1.102	Human	Q02742
β-1,6-GlcNAc transferase IGNT	2.4.1.150	Human	Q06430
β-1,6-GlcNAc transferase C2GNT	2.4.1.102	Mouse	Q09324
Enzymic glycosylation-regulating protein		Rat	S79797
Family 15 (retaining)			
Mannosyltransferase Mnt1	2.4.1.131	*Candida albicans*	X99619
Mannosyltransferase Mnt2	2.4.1.131	*Candida albicans*	P46592
Glycolipid 2-α-mannosyltransferase	2.4.1.131	*Saccharomyces cerevisiae*	P27809
Mannosyltransferase Ktr1	2.4.1.131	*Saccharomyces cerevisiae*	P27810
Mannosyltransferase Ktr2	2.4.1.131	*Saccharomyces cerevisiae*	P33550
Mannosyltransferase Ktr3	2.4.1.131	*Saccharomyces cerevisiae*	P38130
Mannosyltransferase Ktr4	2.4.1.131	*Saccharomyces cerevisiae*	P38131
Mannosyltransferase Ktr5	2.4.1.131	*Saccharomyces cerevisiae*	P53966
Mannosyltransferase Ktr6	2.4.1.131	*Saccharomyces cerevisiae*	P54070
Mannosyltransferase Ktr7	2.4.1.131	*Saccharomyces cerevisiae*	P40504
Mannosyltransferase Yur1	2.4.1.131	*Saccharomyces cerevisiae*	P26725
Family 16 (inverting)			
β-1,2-GlcNAc transferase II	2.4.1.143	Human	Q10469
β-1,2-GlcNAc transferase II	2.4.1.143	Rat	Q09326
Family 17 (inverting)			
β-1,4-GlcNAc transferase III	2.4.1.144	Human	Q09327
β-1,4-GlcNAc transferase III	2.4.1.144	Mouse	Q10470
β-1,4-GlcNAc transferase III	2.4.1.144	Rat	Q02527
Family 18 (inverting)			
β-GlcNAc transferase		*Cricetulus griseus*	U62587
β-GlcNAc transferase	2.4.1.155	Human	Q09328
β-GlcNAc transferase	2.4.1.155	Rat	Q08834
Family 19			
Lipid A disaccharide synthase LpxB	2.4.1.182	*Escherichia coli*	P10441
Lipid A disaccharide synthase LpxB	2.4.1.182	*Haemophilus influenzae*	P45011
LpxB protein		*Proteus mirabilis*	Y09263
LpxB protein		*Synechocystis* sp.	D64000
Family 20 (retaining)			
α,α-Trehalose-phosphate synthase	2.4.1.15	*Arabidopsis thaliana*	Y08568
α,α-Trehalose-phosphate synthase Tps1	2.4.1.15	*Aspergillus niger*	U07184
α,α-Trehalose-phosphate synthase TpsB	2.4.1.15	*Aspergillus niger*	U63416
Unknown ZK54.2		*Caenorhabditis elegans*	U58737
α,α-Trehalose-phosphate synthase Tps1	2.4.1.15	*Candida albicans*	Y07918
α,α-Trehalose-phosphate synthase OtsA	2.4.1.15	*Escherichia coli*	P31677
α,α-Trehalose-phosphate synthase Tps1	2.4.1.15	*Kluyveromyces lactis*	Q07158
OtsA protein		*Mycobacterium leprae*	U15187
α,α-Trehalose-phosphate synthase Tps1	2.4.1.15	*Saccharomyces cerevisiae*	Q00764
α,α-Trehalose-phosphate synthase Tps2	2.4.1.15	*Saccharomyces cerevisiae*	P31688
α,α-Trehalose-phosphate synthase Tps3	2.4.1.15	*Saccharomyces cerevisiae*	P38426
α,α-Trehalose-phosphate synthase Tps1	2.4.1.15	*Schizosaccharomyces pombe*	P40387
OtsA protein		*Synechocystis* sp.	D90913
Family 21 (retaining)			
Unknown F20B4.6		*Caenorhabditis elegans*	U58735
Unknown T06C12.C		*Caenorhabditis elegans*	Z81116
Unknown YK29C8.5		*Caenorhabditis elegans*	U53332
Ceramide glucosyltransferase	2.4.1.80	Human	Q16739
Unknown (ORF SLR0813)		*Synechocystis* sp.	D90911

Table 1 (cont.)

Description[a]	EC number[b]	Organism	Accession no.[c]
Family 22			
Unknown C14A4.3		*Caenorhabditis elegans*	Z49909
Mannosyltransferase		Human	D42138
Mannosyltransferase		*Saccharomyces cerevisiae*	X96417
Family 23 (inverting)			
Fucosyltransferase (NodZ)		*Azorhizobium caulinodans*	L18897
Fucosyltransferase (NodZ)		*Bradyrhizobium japonicum*	L22756
N-Acetyl-β-D-glucosaminide α-1,6-fucosyltransferase		Pig	D86723
Family 24			
Unknown C12C8.D		*Caenorhabditis elegans*	Z81467
UDP-glucose glycoprotein glucosyltransferase		*Drosophila melanogaster*	U20554
Killer toxin-resistance protein Kre5		*Saccharomyces cerevisiae*	P22023
UDP-glucose glycoprotein glucosyltransferase		*Schizosaccharomyces pombe*	U38417
Family 25			
Unknown D2045.9		*Caenorhabditis elegans*	Z35639
LbgA protein		*Haemophilus ducreyi*	U58147
Lic2B protein		*Haemophilus influenzae*	U36398
LPS biosynthesis protein		*Haemophilus influenzae*	X56903
LPS biosynthesis protein		*Haemophilus influenzae*	L19441
LPS biosynthesis protein Lex-1		*Haemophilus influenzae*	U32736
Lex2B protein		*Haemophilus influenzae*	U05670
Galactosyltransferase LgtB		*Neisseria gonorrhoeae*	U14554
Galactosyltransferase LgtE		*Neisseria gonorrhoeae*	U14554
Glycosyltransferase LgtB		*Neisseria meningitidis*	U25839
Glycosyltransferase LgtE		*Neisseria meningitidis*	U25839
LpsA protein		*Pasteurella haemolytica*	U15958
Family 26			
AceB protein		*Acetobacter xylinum*	X94981
UDP-*N*-acetylmannosaminuronic acid transferase		*Escherichia coli*	P27836
Teichoic acid biosynthesis protein A		*Escherichia coli*	P27620
UDP-*N*-acetylmannosaminuronic acid transferase		*Salmonella typhimurium*	P37457
Cps19fF protein		*Streptococcus pneumoniae*	U09239
Unknown (ORF SLR1118)		*Synechocystis* sp.	D90899
Unknown (ORF SLR1271)		*Synechocystis* sp.	D90913
GumM protein		*Xanthomonas campestris*	U22511

described by the present work could provide an aid to structural interpretation and, when more structures become available, suggest possible search models for molecular replacement.

For those families where both the NDP-sugar and the linkage formed are known, i.e., all families except families 19, 22, 24, 25 and 26, the classification based on sequence similarity consistently differentiates retaining from inverting enzymes. This is consistent with the conservation of the catalytic machinery of these enzymes within each family. Almost half of the classified sequences have unknown or uncertain functions (Table 1). A fundamental basis for a classification must be that it has predictive power. The present classification allows the prediction of their global function (i.e. NDP-glycosyltransferase) and product stereochemistry (inverted or retained anomeric configuration).

By analogy to glycoside hydrolases, the catalytic machinery of glycosyltransferases is likely to involve Asp and/or Glu residues whose side chains have the appropriate reactivity to act as the general base for acceptor activation or as the nucleophile for the formation of a glycosyl-enzyme intermediate. Site-directed mutagenesis of ribosyltransferases has shown that specific Glu residues are essential for glycosyltransferase activity [20]. For each family of glycosyltransferases, the list of the invariant Asp or Glu residues is therefore likely to contain catalytic residues. In some of the families we describe there are so few such conserved residues that the catalytic machinery is probably directly identifiable. Examples include families 1, 9 and 11 with one invariant Asp, families 5 and 25 with one invariant Glu, families 3 and 4 with two invariant Glu and families 2, 8, 9 and 20 with two invariant Asp residues.

The EC recommendations place all hexosyltransferases in the same subclass (EC 2.4.1.*x*), regardless of the sugar donor used by the enzymes. There are clear structural, evolutionary and mechanistic similarities between several glycosyltransferases using glycosides as activated sugar donors and glycoside hydrolases. For example, cyclodextrin glucanotransferases (EC 2.4.1.19) and starch branching enzymes (EC 2.4.1.18) are clearly related to a large number of starch-hydrolysing enzymes forming family 13 of glycoside hydrolases [4,6,7,21]. Similarly, endo-xyloglucan transferases (EC 2.4.1.207) display significant similarities to glycoside hydrolase family 16 members [7]. In contrast, we have been unable to detect any sequence similarity between the NDP-sugar glycosyltransferases we have analysed and glycoside hydrolases. This probably reflects particular constraints on the active site of these glycosyltransferases which must accommodate the bulky NDP-moiety.

On several occasions, we observed that enzymes acting on similar substrates with the same mechanism, and classified in different families, displayed intriguing local similarities which could not be extended to the rest of the sequence. This situation, which perhaps reflects the limitations of sequence comparisons at

very high divergence, is reminiscent of the grouping of glycoside hydrolase families into clans where the only sequence similarity is found around the catalytic machinery [22]. An example of such possibly related families are families 3, 4 and 5, which display limited local similarities. Similarly, families 11 and 23 could perhaps be grouped based on the specific instance of the motif VHVRRTD in a family 23 enzyme (porcine *N*-acetyl-β-D-glucosaminide α-1,6-fucosyltransferase) which is almost identical with one of the three highly conserved motifs in family 11 (VHVRRGD motif). Conversely, the proposed grouping of α-1,3-fucosyltransferases and α-1,2-fucosyltransferases [23] cannot be confirmed as the corresponding families (10 and 11) do not bear even one conserved residue. Only structural resolution will allow the reliable grouping of families into 'superfamilies' or 'clans'.

That there are several polyspecific families leads to the proposition that the observed differences in substrate specificity probably reflect divergent evolution from an ancestral form of glycosyltransferase. Conversely, we have identified at least one example of an enzyme activity (lipopolysaccharide 1,2-*N*-acetylglucosaminyltransferase; EC 2.4.1.56) which appears in two distinct families (4 and 9), suggesting that this could constitute an example of convergent evolution.

Genome sequencing projects are increasingly delivering large numbers of potential glycosyltransferase sequences and the present classification that brings together structural, mechanistic and sequence-based information is clearly of biocomputing importance. Significantly, a possible function was recently proposed for secreted Fringe-like signalling molecules based on distant sequence similarity with glycosyltransferases sequences [24]. It is our intention to set up an electronic access to this classification similar to that already implemented for the glycosidases [7].

J. C. thanks the Australian Government Department of Industry, Science and Tourism for financial support of his visit to CERMAV. J.C. also thanks the Sugar Research and Development Corporation for partial funding of this research. G.J.D. is a Royal Society University Research Fellow.

James A. CAMPBELL*, Gideon J. DAVIES†, Vincent BULONE‡§ and Bernard HENRISSAT‡

*CSIRO Tropical Agriculture, 306 Carmody Road, St. Lucia, Q 4067, Australia, †Department of Chemistry, University of York, Heslington, York YO1 5DD, U.K., and ‡Centre de Recherches sur les Macromolécules Végétales (Affiliated with the Joseph Fourier University), CNRS, BP 53, F-38041 Grenoble Cedex 9, France

§ To whom correspondence should be addressed.

1 Kleene, R. and Berger, E. G. (1993) Biochim. Biophys. Acta **1154**, 283–325
2 Sinnott, M. L. (1990) Chem. Rev. **90**, 1171–1202
3 International Union of Biochemistry and Molecular Biology (1992) Enzyme Nomenclature: Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology, Academic Press, San Diego
4 Henrissat, B. (1991) Biochem. J. **280**, 309–316
5 Rawlings, N. D. and Barrett, A. J. (1993) Biochem. J. **290**, 205–218
6 Henrissat, B. and Bairoch, A. (1993) Biochem. J. **293**, 781–788
7 Henrissat, B. and Bairoch, A. (1996) Biochem. J. **316**, 695–696
8 Rawlings, N. D. and Barrett, A. J. (1994) Methods Enzymol. **244**, 1–15
9 Rawlings, N. D. and Barrett, A. J. (1995) Methods Enzymol. **248**, 105–120
10 Davies, G. and Henrissat, B. (1995) Structure **3**, 853–859
11 Gebler, J., Gilkes, N. R., Claeyssens, M., Wilson, D. B., Béguin, P., Wakarchuk, W. W., Kilburn, D. G., Miller, Jr., R. C., Warren, R. A. J. and Withers, S. G. (1992) J. Biol. Chem. **267**, 12559–12561
12 Geremia, R. A., Petroni, A., Ielpi, L. and Henrissat, B. (1996) Biochem. J. **318**, 133–138
13 Saxena, I., Brown, Jr., R. M., Fèvre, M., Geremia, R. A. and Henrissat, B. (1995) J. Bacteriol. **177**, 1419–1424
14 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990) J. Mol. Biol. **215**, 403–410
15 Durand, P., Canard, L. and Mornon, J. P. (1997) Comput. Appl. Biosci., in the press
16 Gaboriaud, C., Bissery, V., Benchetrit, T. and Mornon, J. P. (1987) FEBS Lett. **224**, 149–155
17 Lemesle-Varloot, L., Henrissat, B., Gaboriaud, C., Bissery, V., Morgat, A. and Mornon, J. P. (1990) Biochimie **72**, 555–574
18 Chothia, C. and Lesk, A. M. (1986) EMBO J. **5**, 823–826
19 Vrielink, A., Rüger, W., Driessen, H. P. C. and Freemont, P. S. (1994) EMBO J. **13**, 3413–3422
20 Takada, T., Iida, K. and Moss, J. (1995) J. Biol. Chem. **270**, 541–544
21 Jespersen, H. M., MacGregor, E. A., Henrissat, B., Sierks, M. R. and Svensson, B. (1993) J. Protein Chem. **12**, 791–805
22 Henrissat, B., Callebaut, I., Fabrega, S., Lehn, P., Mornon, J. P. and Davies, G. (1995) Proc. Natl. Acad. Sci. U.S.A. **92**, 7090–7094
23 Breton, C., Oriol, R. and Imberty, A. (1996) Glycobiology **6**, vii–xii
24 Yuan, Y. P., Schultz, J., Mlodzik, M. and Bork, P. (1997) Cell **88**, 9–11

Received 23 April 1997

The Factor I and follistatin domain families: the return of a prodigal son

The Factor I/membrane-attack complex (FIMAC) domain has been identified in three complement proteins of immune defence and was believed to be unique [1]. It is present in the essential regulatory protease Factor I (FI) that is specific for the complement proteins C3b and C4b when complexed with their cofactors [2], and also in the terminal components C6 and C7 of the membrane-attack complex [3,4]. FI contains an N-terminal FIMAC domain (67 residues), a CD5 or scavenger-receptor cysteine-rich domain, two low-density-lipoprotein-receptor domains, and a serine-protease domain. Both C6 and C7 contain two C-terminal FIMAC domains. Here we show that database searches based on the FIMAC sequence indicated similarities with the follistatin (FS) sequence family (Figure 1). Similarities were also found in the predicted secondary structures of the FIMAC and FS domains (Figure 2). We therefore propose that the FIMAC domain is a distant member of the larger FS superfamily [5,6]. For FI, all the domain types in its structure have now been identified as members of larger cellular-receptor or extracellular-matrix domain superfamilies [7–9]. The FS superfamily of the extracellular matrix and the endocrine system now includes the immune equivalent of this domain.

Seven FIMAC sequences [3,4,10–12] were used in database searches of more than 59000 SWISSPROT sequences using BLITZ which is based on the program MPsrch with the BLOSUM62 amino acid substitution matrix [13]. Table 1 indicates hits with agrin in the FS family with 19–24% sequence matches. Lower matches in the top 50 hits (Table 1) included

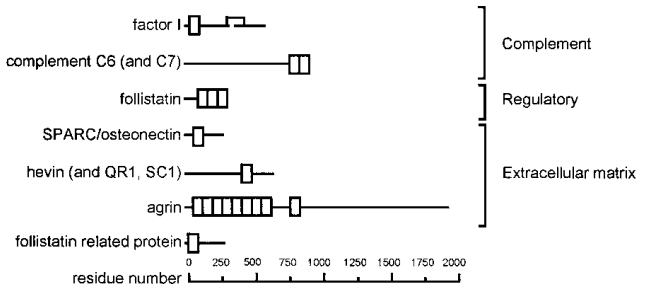

Figure 1 Occurrence of FIMAC and FS domains in the sequences of plasma, regulatory and extracellular matrix proteins

Each FIMAC or FS domain is denoted by a box.

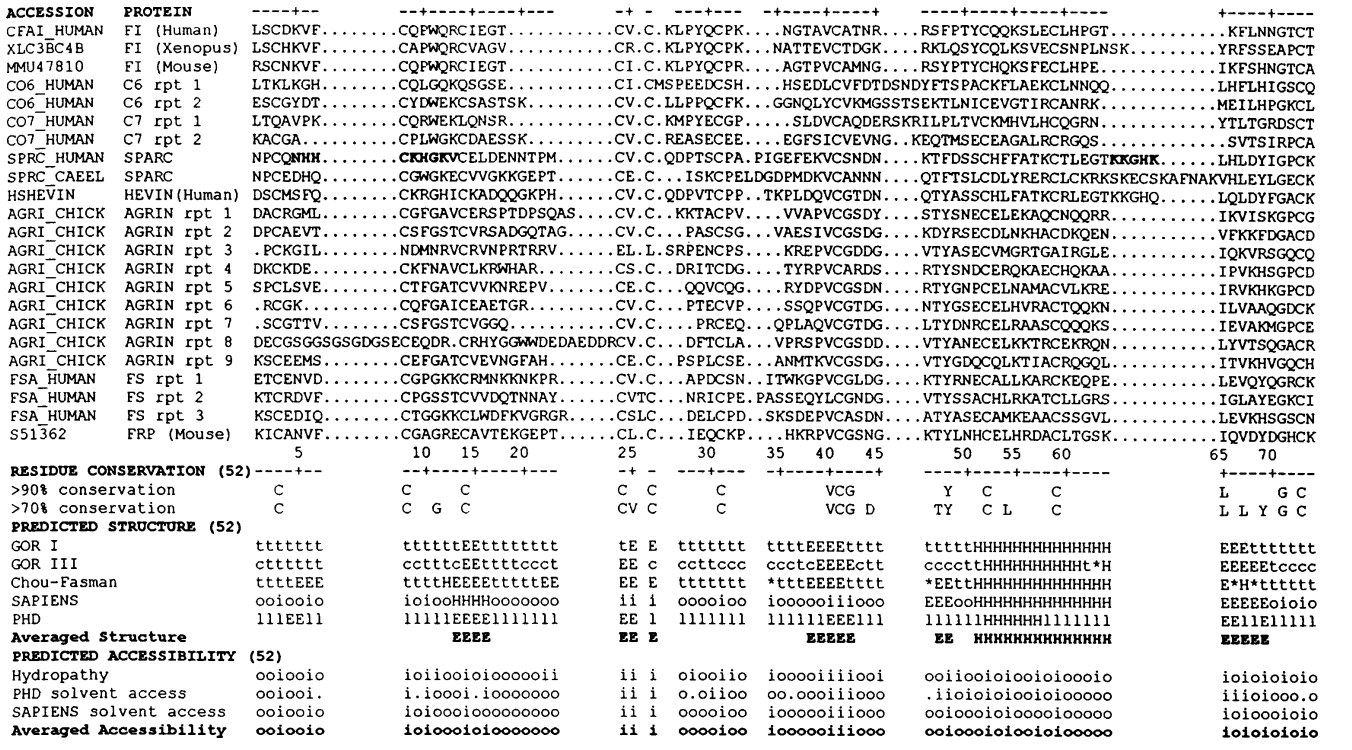

Figure 2 Summary of the multiple sequence alignment of FIMAC/FS sequences

A total of 52 was used in the full alignment. The 29 not shown in the Figure have SWISSPROT accession codes SPRC_BOVIN, SPRC_MOUSE, SPRC_RAT, SPRC_CHICK, SPRC_XENLA, QR1_COTJA, SC1_RAT, DYGAGR, AGRI_RAT, FSA_PIG, FSA_RAT, FSA_SHEEP, FSA_XENLA, from which the full alignment is readily reconstructed. The two Cu^{2+}-binding regions of the SPARC sequences are in **bold**. The alignment yielded a consensus length of 74 residues which is conserved in over 50% of the 52 sequences. The most commonly occurring residues that show 90% and 70% conservation within a given subtype (I=V=L=M; D=E; R=K=H; F=Y=W=H; G=A=S) are indicated. The averaged predicted structures are based on the presence of at least 2/5 secondary structures or 2/3 accessibility states at each residue position. Abbreviations are as follows; H, α-helix; E, β-sheet; l, c and t, loop, coil and turn; o, solvent-accessible; i, solvent-inaccessible.

Table 1 BLITZ database search of more than 59000 sequences

Probe FIMAC sequence (see Figure 2)	FS sequences detected (P better than 10^{-4})	Probability (P)	Total of FS sequences in top 50 positions
CFAI_HUMAN (residues 41–107)	AGRI_RAT	4×10^{-5}	10
XLC3BC4B (residues 39–101)	AGRI_RAT	4×10^{-6}	9
MMU47810 (residues 44–110)	AGRI_CHICK	3×10^{-7}	16
	AGRI_RAT	9×10^{-7}	
CO6_HUMAN (residues 766–837)	None		17
CO7_HUMAN (residues 771–839)	AGRI_CHICK	7×10^{-8}	16
	AGRI_RAT	9×10^{-6}	

other agrin repeats, FS, SPARC (**s**ecreted **p**rotein, **a**cidic, **r**ich in **c**ysteine; also known as osteonectin) and the SPARC homologues QR1 and SC1, all of which belong to the FS family. The BLITZ matches with the FS domains of chicken and rat agrin extended over 52–65 residues of the probe FIMAC sequences (Table 1) and aligned eight to ten of the eight or ten cysteine residues in the FIMAC sequences. Another database search using BLASTP [14] to scan more than 71000 sequences showed that AGRI_RAT and AGRI_CHICK were scored with statistically significant probabilities (P) of 10^{-5}–10^{-6} by FIMAC sequences from FI and C7. Likewise version 3 of FASTA [15] yielded similar results with more than 59000 sequences, in which FS domains in agrin, QR1 and FS also scored highly with probabilities E(59000) less than 0.05, and lower matches included other FS domains. FASTA matched eight to ten of the eight or ten cysteine residues of FIMAC with their equivalents in the FS domain. Blockmaker [16], with the FIMAC sequences and the FS domains from agrin and SPARC (Figure 2), produced two blocks, one of which aligned with both the FIMAC and FS sequences. Likewise, MACAW [17] generated a 28-residue block of ten aligned sequences from the top matches in the BLITZ search, which corresponded to that identified by Blockmaker [16] (residues 31–60 in Figure 2). Whereas the 28-residue block lacked the C6 or C7 FIMAC sequences, its occurrence was statistically significant.

Since these analyses showed a relationship between the FIMAC and FS sequences, a combined alignment was constructed using 52 sequences (Figure 2). The consensus length is 74 residues with ten conserved cysteine residues. These cysteine residues are assumed to be bridged. Although the disulphide pairings are unknown, cysteine mutations in the first FIMAC of C6 and C7 suggest that Cys^3–Cys^{14} are paired (the first and third Cys in

Figure 2). Cys^{8}, Cys^{24}, Cys^{26} and Cys^{59} (the second, fourth, fifth and ninth Cys in Figure 2) are missing in agrin domain 3 and may form two pairs if the domain structure is unaltered. Whereas C-terminal sequence similarities with the disulphide-rich Kazal-type inhibitors of the ovomucoid family and N-terminal similarities with the epidermal-growth-factor family have been noted [6,8], the proposed disulphide pairings in the alignment are mutually exclusive in the two families [5]. Figure 2 also shows conserved buried hydrophobic residues. These are attributable to the packing of the protein core, and most are present in both the FIMAC and FS sequences.

The FIMAC and FS sequences were also compared by computing consensus secondary-structure predictions from the alignment [18,19]. Use of the GORI, GORIII, Chou Fasman, PHD and SAPIENS prediction algorithms gave an averaged $\beta\beta\beta\beta\alpha\beta$ structure with five β-strands and one α-helix (Figure 2). Residues 38–74 resemble the Kazal-type inhibitors. Interestingly, they resulted in a $\beta\beta\alpha\beta$ prediction that is very similar to the observed $\beta\beta\alpha\beta$ secondary structure in ovomucoid when analysed using DSSP (Brookhaven database codes: 1ovo-4ovo). Cys^{35} and Cys^{38} are located on the α-helix of ovomucoid and correlate well with Cys^{52} and Cys^{59} in FIMAC/FS, which are located on the α-helix predicted between residues 51 and 64 (Figure 2). Cys^{52} and Cys^{59} in FIMAC/FS will be positioned on the same side of this α-helix as Cys^{35} and Cys^{38} in ovomucoid, since Cys^{52} and Cys^{59} will be separated by an extra turn of the predicted α-helix in the FIMAC/FS domain compared with ovomucoid. The presence of the extra N-terminal residues in FIMAC before the region of similarity with ovomucoid implies that a structural relationship to ovomucoid is possible, but this will be modified. Figure 2 also shows large sequence insertions in the alignment, and these occur in regions predicted to be surface loops as desired. The loop between residues 64 and 65 corresponds to the Cu^{2+}-binding region of the SPARC proteins [5] that is implicated in cellular proliferation, but is absent from other FIMAC and FS domains. More importantly, application of averaged secondary-structure predictions to each of the FIMAC and FS sequence families yielded results that were very similar to the predicted $\beta\beta\beta\beta\alpha\beta$ structure for all 52 sequences and support the proposed identity between the two families.

The phylogenic relationship between the FIMAC and FS sequences was investigated using PHYLIP [20]. An unrooted tree showed that the FIMAC sequences occupied a separate branch from the FS sequences. The lengths of the exons in SPARC, agrin, FI, C6 and C7 are in agreement with Figure 2, whereas their boundaries are not conserved. In murine SPARC, the FS domain is encoded by exons 5 and 6 with intron boundaries of class 1-1 [5], and correlates well with the FS domains in agrin, which are encoded by one or two exons and have intron boundaries of class 1-1 [21]. In human FI, the FIMAC domain is within exon 2, but with intron boundaries of class 0-1 [22]. The two FIMAC domains in human C6 and C7 are encoded across exons 15, 16 and 17 with intron boundaries of class 1-2, 2-1 and 1-undefined respectively [3,4]. The lack of conserved intron/exon structure is similar to that found in the serine-protease domain [22].

The relationship between the FIMAC and FS families is matched functionally in that all proteins containing FIMAC and FS domains are extracellular and participate in protein-protein interactions (Figure 1). The FS domain has been identified in extracellular matrix proteins that modulate cell–matrix interactions (SPARC), induce aggregation of nicotinic acetylcholine receptors (agrin) and in ovaries and the pituitary that bind cytokines (FS) [5,6]. Hevin is isolated from high endothelial venules of tonsils which allow high levels of lymphocyte extravasion from blood and may facilitate this lymphocyte migration [23]. SPARC is released by platelet degranulation, is synthesized by fibroblasts and macrophages at sites of wound repair, and may regulate deposition or assembly of extracellular matrix proteins. The specificity of the FS domain for its ligand is indicated by SPARC and FS. SPARC binds to platelet-derived growth factor, albumin, thrombospondin and various collagen types. FS binds to activin and inhibin, which are transforming-growth-factor-β-like cytokines. Despite the sequence similarity to ovomucoid, no protease-inhibitory activity has been reported to date for FS-containing proteins [6]. FI interacts with the complement components C3b and C4b, whereas C6 and C7 interact with complement component C5b during formation of the membrane-attack complex. As C3b, C4b and C5b are all related in sequence, it will be of interest to determine whether all three contain a similar target fold for FIMAC. Given the relationship to ovomucoid, it will be of interest to determine whether FIMAC in FI can inhibit its own serine-protease domain. From X-ray and neutron-scattering analyses of FI, one model that is consistent with the data is a semi-compact V-shaped structure, the dimensions of which place these two domains proximate to each other [24].

We thank the Wellcome Trust for grant support and Dr. R. B. Sim for useful discussions.

Christopher G. ULLMAN and Stephen J. PERKINS[1]

Department of Biochemistry and Molecular Biology, Royal Free Hospital School of Medicine, Rowland Hill Street, London NW3 2PF, U.K.

[1] To whom correspondence and requests for offprints should be addressed.

1 Law, S. K. A. and Reid, K. B. M. (1995) Complement, 2nd edn., IRL Press, Oxford
2 Sim, R. B., Day, A. J., Moffatt, B. E. and Fontaine, M. (1993) Methods Enzymol. **223**, 13–35
3 Hobart, M. J., Fernie, B. and DiScipio, R. G. (1993) Biochemistry **32**, 6198–6205
4 Hobart, M. J., Fernie, B. A. and DiScipio, R. G. (1995) J. Immunol. **154**, 5188–5194
5 Lane, T. F. and Sage, E. H. (1994) FASEB J. **8**, 163–173
6 Patthy, L. and Nikolics, K. (1993) Trends Neurosci. **16**, 76–81
7 Perkins, S. J. and Smith, K. F. (1993) Biochem. J. **295**, 109–114
8 Moestrup, S. K. (1994) Biochim. Biophys. Acta **1197**, 197–213
9 Resnick, D., Pearson, A. and Krieger, M. (1994) Trends Biochem. Sci. **19**, 5–8
10 Catterall, C. F., Lyons, A., Sim, R. B., Day, A. J. and Harris, T. J. (1987) Biochem. J. **242**, 849–856
11 Kunnath-Muglia, L. M., Chang, G. H., Sim, R. B., Day, A. J. and Ezekowitz, R. A. (1993) Mol. Immunol. **30**, 1249–1256
12 Minta, J. O., Wong, M. J., Kozak, C. A., Kunnath-Muglia, L. M. and Goldberger, G. (1996) Mol. Immunol. **33**, 101–112
13 Sturrock, S. S. and Collins, J. F. (1993) MPsrch, Version 1.5, Biocomputing Research Unit, University of Edinburgh
14 Altschul, S. F., Warren, G., Miller, W., Myers, E. W. and Lipman, D. J. (1990) J. Mol. Biol. **215**, 403–410
15 Pearson, W. R. and Lipman, D. J. (1988) Proc. Natl. Acad. Sci. U.S.A. **85**, 2444–2448
16 Henikoff, S., Henikoff, J. G., Alford, W. J. and Pietrokovski, S. (1995) Gene **163**, GC (Gene-COMBIS) 17–26
17 Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F. and Wootton, J. C. (1993) Science **262**, 208–214
18 Edwards, Y. J. K. and Perkins, S. J. (1996) J. Mol. Biol. **260**, 277–285
19 Brissett, N. C. and Perkins, S. J. (1996) FEBS Lett. **388**, 211–216
20 Felsenstein, J. (1989) Cladistics **5**, 164–166
21 Rupp, F., Özçelik, T., Linial, M., Peterson, K., Francke, U. and Scheller, R. (1992) J. Neurosci. **12**, 3535–3544
22 Vyse, T. J., Bates, G. P., Walport, M. J. and Morley, B. J. (1994) Genomics **24**, 90–98
23 Girard, J. P. and Springer, T. A. (1995) Immunity **2**, 113–123
24 Perkins, S. J., Smith, K. F. and Sim, R. B. (1993) Biochem. J. **295**, 101–108

Received 6 May 1997