Bacterial polysaccharide synthesis and gene nomenclature

Peter R. Reeves, Matthew Hobbs, Miguel A. Valvano, Mikael Skurnik, Chris Whitfield, David Coplin, Nobuo Kido, John Klena, Duncan Maskell, Christian R.H. Raetz and Paul D. Rick

Gene nomenclature for bacterial surface polysaccharides is complicated by the large number of structures and genes. We propose a scheme applicable to all species that distinguishes different classes of genes, provides a single name for all genes of a given function and greatly facilitates comparative studies.

P.R. Reeves* and M. Hobbs are in the Dept of Microbiology, University of Sydney, NSW 2006, Australia; M.A. Valvano is in the Dept of Microbiology and Immunology, University of Western Ontario, London, Ontario, Canada N6A 5C1; M. Skurnik is in the Turku Centre for Biotechnology, PO Box 123, 20521 Turku, Finland; C. Whitfield is in the Dept of Microbiology, University of Guelph, Guelph, Ontario, Canada N1G 2W1; D. Coplin is in the Dept of Plant Pathology, Ohio State University, Columbus, OH 43210-1087, USA; N. Kido is in Biosystems, School of Informatics & Sciences, Nagoya University, Nagoya 464-01, Japan; J. Klena is in the Dept of Plant and Microbial Sciences, University of Canterbury, Christchurch 4, New Zealand; D. Maskell is in the Dept of Clinical Veterinary Medicine, University of Cambridge, Cambridge, UK CB3 0ES; C.R.H. Raetz is in the Dept of Biochemistry, Duke University Medical Center, Durham, NC 27710, USA; P.D. Rick is in the Dept of Microbiology and Immunology, The Uniformed Services University of the Health Sciences, Bethesda, MD 20814, USA.
*tel: +61 2 9351 2536, fax: +61 2 9351 4571, e-mail: reeves@angis.su.oz.au

Most bacteria produce surface and/or secreted polysaccharides that can act as prominent antigens. Many of these polysaccharides are also extremely variable, as shown for *Salmonella* and *Escherichia coli*. For a review of *E. coli* serology see Ref. 1, and for reviews of the biochemical and genetic basis of the variation see Refs 2–6.

Diversity within *Salmonella* is equivalent to that within a species such as *E. coli*, and many researchers place all *Salmonella* within *S. enterica*. However, the nomenclature in common use still reflects the historical allocation of species names to each serovar, for example *S. typhimurium*. In the context of this review, we will use the genus name *Salmonella* where possible. On the basis of DNA relatedness, *E. coli* and the four species of *Shigella* are a single species, and therefore we will treat strains of *Shigella* as serovars within the species *E. coli*, for example *E. coli* dysenteriae.

Bacterial polysaccharides include lipopolysaccharide (LPS), lipooligosaccharide (LOS) and extracellular polysaccharide (EPS) (Fig. 1). LPS is present in most Gram-negative bacteria and, characteristically, comprises three components: lipid A, core oligosaccharide and O antigen. The lipid A component is composed of sugars and fatty acids, which anchor the LPS in the outer leaflet of the outer membrane and where, in some species at least, it replaces phospholipid. The core is made of sugars and sugar derivatives, such as 3-deoxy-D-manno-octulosonic acid (Kdo). The O antigen is a polysaccharide that extends from the cell surface and consists of repeating oligosaccharide units generally composed of 3–6 sugars (O units; often repeated 10–30-fold). LOS lacks a polymeric O antigen.

EPS may be present in both Gram-negative and Gram-positive bacteria and, like O antigen, is made of repeating units. EPS can be released from the cell into the environment as a slime or remain attached to the cell to form a capsule. The distinction between LPS and different forms of EPS is not always clear and the allocation of names is sometimes close to arbitrary[6].

Variation within and between species

In the majority of the species studied, the O antigen has been found to be highly polymorphic and in general even related species have few or no O-antigen types in common. For example, of the approximately 50 and 173 O antigens found in *Salmonella* and *E. coli*, respectively, only three are common to both species. Similarly, there are few O antigens common to *E. coli* and *Klebsiella*. The total number of types present is still likely to be many thousands. Likewise, there are 80 capsule types in *E. coli*[7]. Many other species have a large repertoire of O antigens and capsules. Capsules also occur in many Gram-positive species, and EPS are also widespread.

Kenne and Lindberg[8] list 105 sugars found in bacterial polysaccharides and 233 structures. Many other polysaccharides have been identified antigenically but not characterized chemically. The variation is caused not only by the diversity of monosaccharide components, but also by the diversity of linkages between sugars. Further levels of variation are imparted by the addition of non-sugar moieties (such as O-acetyl residues or amino acids) and variation in the modal length of the polysaccharide chain. A comprehensive list would be extremely long; readers may wish to refer to the complex carbohydrate structural database (CCSD)[9].

0966 842X/96/$15.00 PII: S0966-842X(96)10071-8

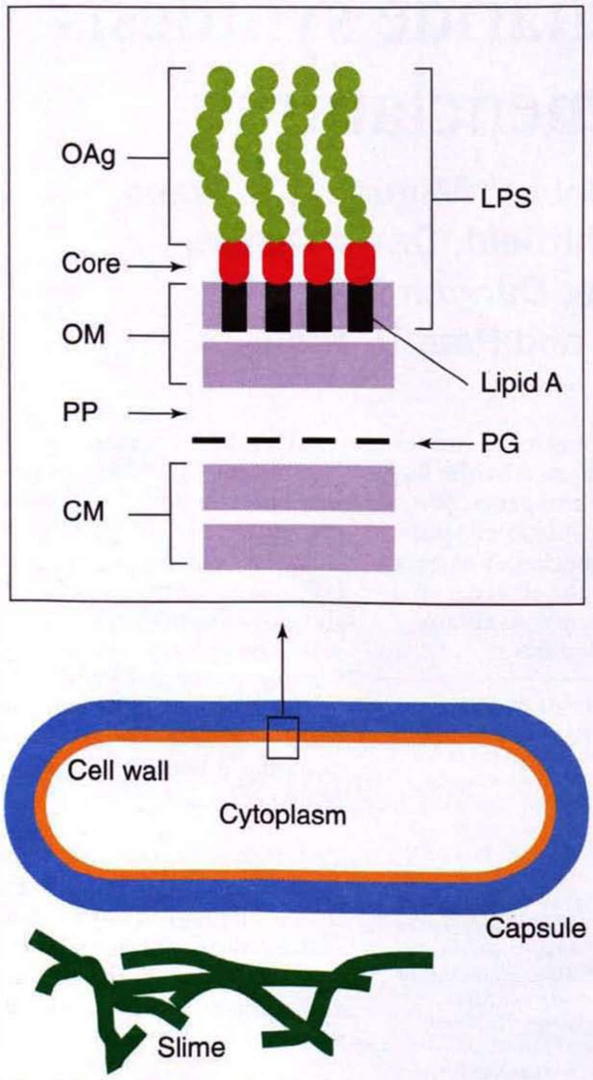

Fig. 1. Bacterial polysaccharides. Lipopolysaccharide (LPS) is present in most Gram-negative bacteria and comprises lipid A (black), core oligosaccharide (red) and O antigen (OAg; light green). Extracellular polysaccharides (EPS) can be released from the cell into the environment as a slime (dark green) or remain attached to the cell to form a capsule (blue). The bacterial cell wall is depicted in orange. Abbreviations: OM, outer membrane; PP, periplasm; PG, peptidoglycan; CM, cytoplasmic membrane.

Current nomenclature for polysaccharide genes

The Demerec system of genetic nomenclature for bacteria[10] has four letters per gene with the last being capitalized. This system continues to work well in most cases but the diversity of bacterial polysaccharides and, hence, of genes for their synthesis poses major problems. Genes involved in the biosynthesis of surface polysaccharides are generally arranged in clusters. *E. coli* has several such clusters, including *rfa* (LPS core), *rfb* (O antigen), *cps* (group I capsules and colanic acid), *rfe–rff* (enterobacterial common antigen; ECA) and *kps* (group II K capsules). These gene clusters have, in general, three classes of genes: those required for the enzymes involved in the biosynthetic pathways of nucleotide sugars, or other components, needed for polysaccharide synthesis and not otherwise available in the cell; genes for the glycosyl transferases and genes for oligosaccharide or polysaccharide processing. Depending on the complexity of the saccharide, the number of genes in the clusters varies from ~6–19 but is not likely to exceed 26 (the number of letters in the alphabet) in any cluster. However, each O antigen, capsule, EPS or LPS core type has its own set of genes.

There are already many more than 26 different genes known for O-antigen synthesis in *Salmonella* alone. The total number of O-antigen genes for *Salmonella* or *E. coli* will be much greater than 26, and the combined total will be greater still because these bacteria have few O antigens in common and many *rfb* clusters will have at least one transferase gene unique to that cluster. Use of the *rfb* symbol set has also been extended to other genera (*Klebsiella*, *Vibrio*, *Yersinia*, *Neisseria* and *Xanthomonas*), which, while desirable in one sense, compounds the problem. We point to just one recent case of confusion caused by reuse of names. A gene in the *Mycoplasma genitalium* genome sequence, which shows homology with *rfbD* of *Klebsiella pneumoniae* O1, was assumed to encode a dTDP-4-dehydrorhamnose reductase[11], presumably because of a similar function in the original *rfbD* gene of *E. coli* K-12. However, the *rfbD* gene of *K. pneumoniae* is not related to the *rfbD* gene of *E. coli* K-12, and the assignment was based on a misunderstanding caused by the confusing nomenclature.

Clearly, we have run out of letters for the fourth letter of the *rfb* gene symbol, and many of the *rfb* symbols (such as *rfbA*) have been used to name a variety of different genes. If separate names are to be used for novel O-antigen genes as they are discovered, then the number of symbols available needs to be expanded. Similar problems apply for other gene clusters. Capsule genes have been named *cps* in some species, *cap* in *Staphylococcus aureus*, and both *cap* and *cps* for different capsule types of *Streptococcus pneumoniae*. EPS genes have been named *exo*, *eps* or *gum*, or named after the species, as in the *ams* gene involved in *Erwinia amylovora* amylovoran synthesis. This is further complicated because the *Erwinia stewartii* capsule *cps* gene cluster is very similar to the *E. amylovora* EPS *ams* cluster. Both polysaccharides are now thought of as secreted EPS, but *E. stewartii* retains 'capsule' gene names.

It is desirable for a given gene symbol to have the same meaning in all strains within a species, and there seems considerable merit in making it possible to use the same set of symbols in any of the large number of bacterial species. We put forward a proposal that would make this possible.

A proposal for renaming bacterial surface polysaccharide genes

In the case of polymorphic loci, such as *rfb* and *cps*, the Demerec nomenclature allows only 26 symbols to define the variety of nucleotide sugar biosynthetic genes, transferase genes and other genes because only the fourth letter of the gene designation is used. By also varying the letter in the second and third positions, a total of 17576 symbols becomes available. Thus, we recommend that the use of *rfb*, *cps*, *exo*, *eps*, *cap*, *ams*, *kps* and similar gene names be abandoned, and we propose a new scheme for a bacterial polysaccharide gene nomenclature (BPGN) as follows:

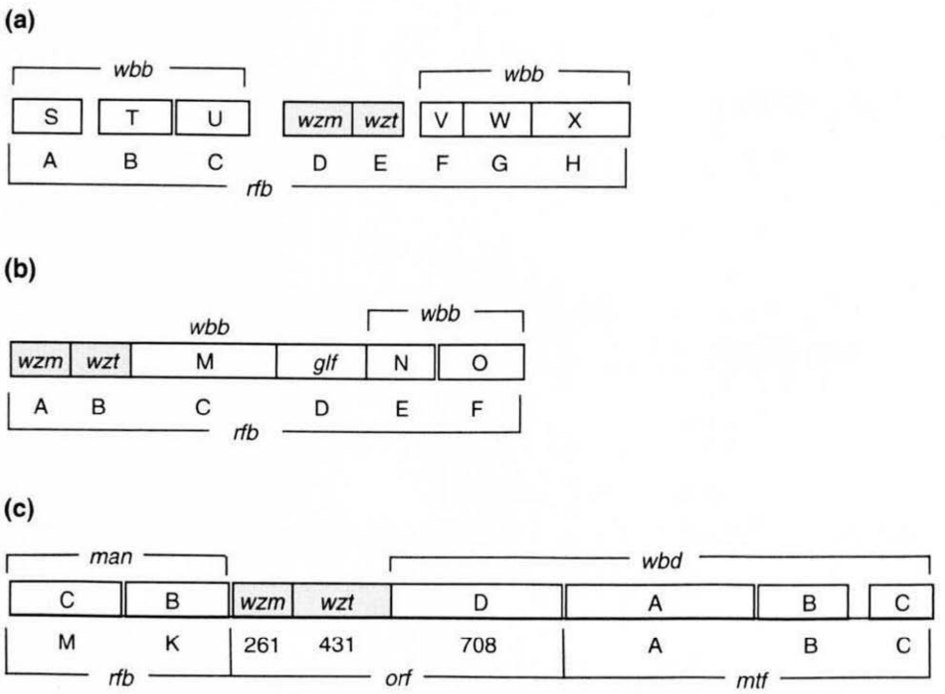

Fig. 2. The O-antigen gene clusters of **(a)** *Yersinia enterocolitica* O3 (Ref. 19), **(b)** *Klebsiella pneumoniae* O1 (Ref. 17) and O8 (R.F. Kelly and C. Whitfield, unpublished) and *Serratia marcescens* O16 (Ref. 18) and **(c)** *Escherichia coli* O9 (Ref. 15). The new nomenclature is used in the boxes (which represent genes) and, for groups of genes, the three-letter names are given above the boxes. Pathway and saccharide-processing genes are shaded. The old nomenclature is given below the boxes.

(1) Most genes should be given names in the form of *w****, which provides 17576 gene names. All genes of any cluster should, in general, have the same first three letters, except where the function is the same as that of a gene in another cluster. In this case, the same name is used for both genes even if it means that genes with different three-letter names coexist in one cluster. The *w*** set was chosen, rather than any other set, because there are very few genes already using this group of symbols. It is proposed that all genes of any block defined by the first two letters are of the same general type. For example, all *wb** genes (*wbaA* to *wbzZ*) will be O-antigen genes. Likewise *wc**, *we** and *wl** can be used for capsule, exopolysaccharide and LOS genes, respectively. By choosing these names we have aimed to maintain a link between current and proposed usage. For example, *wbdA* replaces *mtfA* (Fig. 2) and *wbaP* replaces *rfbP* (Fig. 3). Many other examples are given in Figs 2 and 3. A full list of the gene names allocated at the time of submission is shown in Table 1.

(2) Genes for certain families of homologous proteins involved in saccharide processing, which are common to many gene clusters, are given names of the form *wz**.

(3) Genes involved in the synthesis of saccharide precursors, mostly nucleotide sugars, have names related to the pathway, which are applied to homologous genes in all gene clusters in all species (see Fig. 4).

Ideas concerning the implementation of this scheme can be found in Box 1.

Pathway genes

Some of the sugars found in bacterial polysaccharides are present elsewhere in the cell and their biosynthetic pathways are part of general metabolism, but for the remaining sugars, the genes for biosynthesis will generally be in the gene cluster for the polysaccharide. Most of the sugars occur in more than one structure and many of the pathways have steps in common. As there are relatively few genes in each pathway and each pathway has its own three-letter symbol, the total number of genes involved should be low. The use of names related to the pathway should make it easier to associate genes and functions and to identify homologous genes. For example, genes for the dTDP-L-rhamnose pathway will be named *rmlA–D*. This approach has already been used for pathway genes located within region II of the *kps* cluster, as well as for individual genes such as *galE* (involved in the synthesis of UDP-galactose), which has been found within polysaccharide gene clusters.

Transferase genes

In contrast to the number of pathway genes anticipated, the number of possible linkages, and hence of specific transferases, is very large indeed, and many gene symbols will be required if each unique function is to have a unique name. It is largely because of the number of specific transferases that we chose to use the *w**** set of gene names for bacterial polysaccharide gene names. Each cluster will, in general, have one such gene for each linkage in the structure, although bifunctional transferases are known.

A given gene symbol is used for genes carrying out the same function and/or displaying obvious homology. This principle can be illustrated by some of the glycosyl transferases of *Salmonella* (see Fig. 3). All *wbaP* (*rfbP*) genes encode an enzyme that transfers Gal-1-P to undecaprenol-P in the initiating step of some O units, and all have sequences that are readily aligned. The gene *wbaV* (*rfbV*) from group D and group B strains encodes a tyvelose transferase and an abequose transferase, respectively, but each can carry out both functions in the presence of appropriate precursors. As they display evident homology they are both named *wbaV*; future studies may show that they have a considerable degree of specificity, but that they were

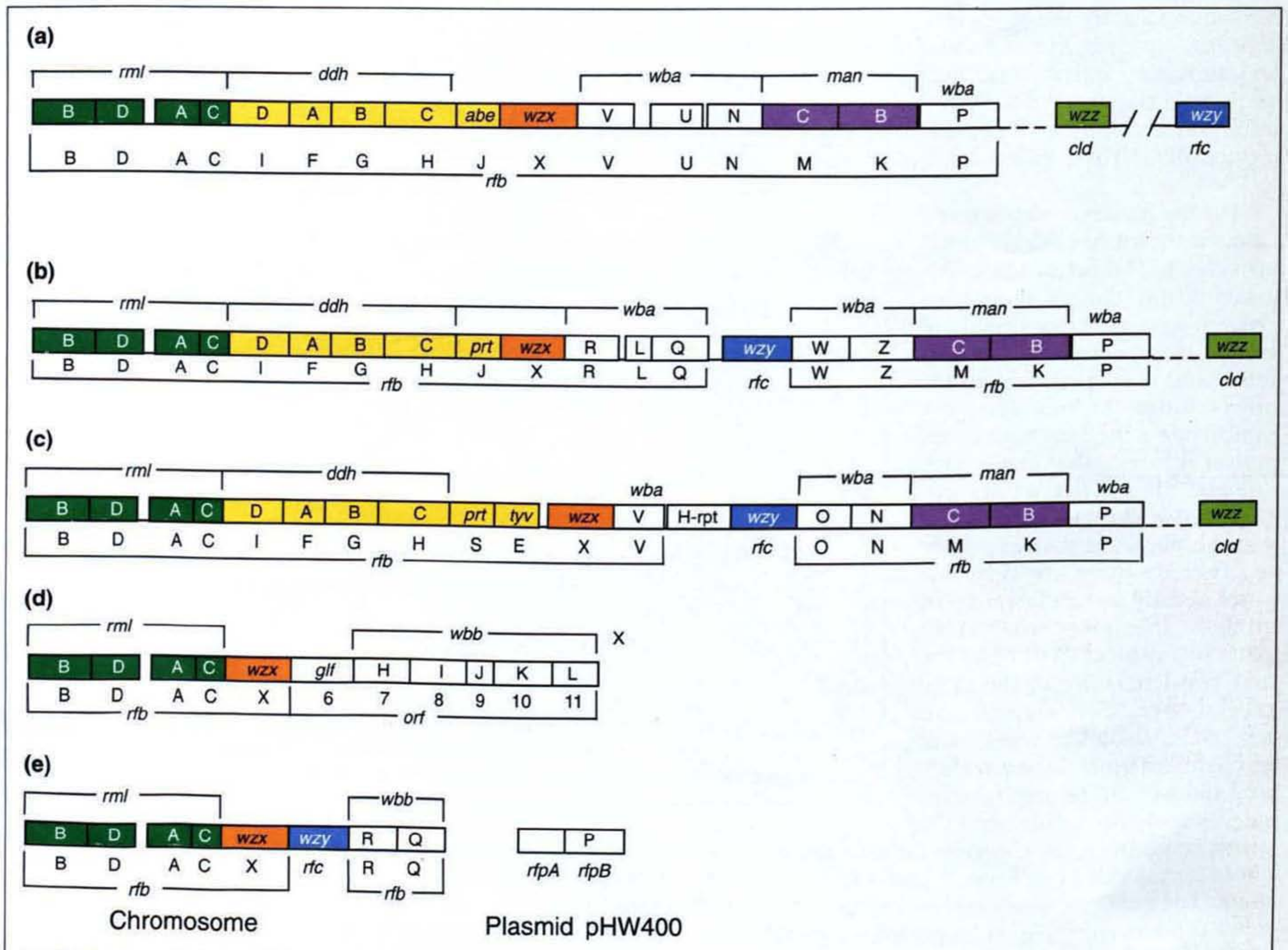

Fig. 3. The gene clusters of *Salmonella* **(a)** group B (Refs 12,26,29), **(b)** group C2 (Refs 12,29,30), **(c)** group D2 (Ref. 31), **(d)** *Escherichia coli* K-12 (016)[32] and **(e)** *E. coli* dysenteriae (type 1)[33]. The nomenclature is as for Fig. 2, and pathway and saccharide-processing genes are shown with pathway-specific colours.

simply not distinguished on their ability to complement *in vivo* in antigen synthesis. Thus, they are given the same name for now. Finally, the 4-mannosyl transferases, which have been shown to form different linkages [Man-(β1-4)-Rha for *wbaO* (*rfbO*), Man-(α1-4)-Rha for *wbaU* (*rfbU*); Man-(α1-2)-Man for *wbaW* (*rfbW*) and Man-(α1-3)-Gal for *wbaZ* (*rfbZ*)] and have barely detectable sequence similarity for the four genes[12], are each given their own names.

Saccharide-processing genes

The genes for saccharide processing (including export, polymerization and assembly of complex polysaccharides such as LPS) commonly occur in families of homologous genes that perform the same general function. The proposal is that all genes of a family are given the same name and that it is in the *wz** subset of *w**** names. This will enable ready identification of such genes. Examples are given in Figs 2 and 3.

Two genes with the same name may have specificity for different oligosaccharides and need to be distinguished by including species and/or other relevant information as a subscript (see below).

Gene clusters containing wzy *(repeat unit polymerase),* wzx *and* wzz *genes*

Most O-antigen gene clusters have genes currently known as *rfc*, *rfbX* and *cld* or *rol*. Each can be recognized by topological features of the encoded protein, although there is usually little sequence similarity for the proteins in the first two families. *wzy* genes from different sources have different specificity and the same may apply to *wzx* genes. The pre-existing names of each gene (*rfc*, *rfbX* and *cld*/*rol*) are anomalous in some way and we suggest that they be renamed *wzy*, *wzx* and *wzz*, respectively, to be consistent with the nomenclature for other genes.

The *wzy* genes are generally found within the O-antigen gene cluster but were not given *rfb** names for historical reasons, as in the classical *Salmonella* strain LT2 the gene maps separately. As we propose to abandon the use of *rfb**, *rfbX* now seems anomalous, and in the case of *rol* and *cld* we have two names for the same gene. A further reason for adopting new terminology is that as a *wzx* gene has been reported for a capsule gene cluster[13] and, because *wzy* may also be found in capsule gene clusters, *rfbX* and *rfc* are inappropriate names.

Table 1. Gene name allocations[a]

3-letter gene names	4th letters of gene names	General description of genes[b,c]
waa	*A–C,E–G,I–N,P,Q,S,Y,Z*	*Escherichia coli* K-12 and *Salmonella* LT2 lipid A/core (*waaA* was *kdtA*, *waaM* and *waaN* were *htrB* and *msbB*, respectively) *waaA and waaC* also in *Bordetella pertussis*
wba	*A–D*	*Salmonella* group C1 O antigen
wba	*E,L,N–R,U–W,Z*	*Salmonella* groups A,B,C2,D1,D2,E O antigens in various combinations
wbb	*D*	*E. coli* O7 O antigen
wbb	*E,F*	*Salmonella* group O54 serovar Borreze
wbb	*H–L*	*E. coli* O16 (K-12) O antigen
wbb	*M–O*	*Klebsiella pneumoniae* O1 and O8 and *Serratia marcescens* O16 O antigens
wbb	*P–R*	*E. coli* dysenteriae type I O antigen
wbb	*S–X*	*Yersinia enterocolitica* O3 O antigen
wbc	*A–J*	*Y. enterocolitica* O8 O antigen
wbc	*K–Q*	*Y. enterocolitica* O3 outer core
wbd	*A–D*	*E. coli* O9 O antigen
wbd	*J,K*	*E. coli* O111 O antigen
wby	*A–E*	*Yersinia pseudotuberculosis* O antigens (incomplete information for several serovars)
wca	*A–M*	*E. coli* K-12 colanic acid
wce	*B,F,G,J–N*	*Erwinia stewartii* stewartan
wec	*A–G*	*E. coli* and *Salmonella* ECA (*rff* and *rfe* genes)
wlb	*A–L*	*B. pertussis* LPS
wza		Homologous genes for OM proteins with a signal peptidase site
wzb		Homologous genes for proteins with acid-phosphatase motif
wzc		Homologous genes for proteins with ATP-binding motif
wzm		Homologues of *kpsM*
wzt		Homologues of *kpsT*
wzx		Replaces *rfbX*
wzy		Replaces *rfc*
wzz		Replaces *rol* and *cld*
abe		CDP–abequose synthesis
alt	*A–C*	dTDP–6-deoxyaltrose pathway
asc	*E,F*	CDP–ascarylose pathway
col		GDP–colitose pathway
ddh	*A–D*	CDP–dideoxyhexose pathway
fcl	*A*, etc.	GDP–L-fucose pathway (*fuc* already used in LT2 and K-12)
fcn	*A,B*	dTDP–fucosamine and dTDP–*N*-acetyl-fucosamine pathway
glf		UDP–galactofuranose synthesis
gmd		GDP–4-keto-6-deoxy-mannose dehydratase
gmh	*A–D*	ADP–glyceromannoheptose pathway
kds	*A,B*	CMP–Kdo pathway
man	*B,C*	GDP–mannose pathway
mna	*A,B*	UDP–*N*-acetyl-D-mannosamine and UDP–*N*-acetyl-D-mannosaminuronic acid pathway
prt		CDP–paratose synthesis
per	*A*, etc.	GDP–perosamine pathway
rml	*A–D*	dTDP–L-rhamnose pathway
rmd		GDP–D-rhamnose synthesis
tyv		CDP–tyvelose synthesis
vsn	*A,B*	dTDP–viosamine pathway (*E. coli* O7, previously misnamed quinovosamine)

[a]Some pathway names have not yet been used (e.g. *per*, *col* and *fcl*) but are reserved for the uses shown in Fig. 4. References can be found in the bacterial polysaccharide gene nomenclature (BPGD; see Box 2).
[b]Description of *wz** genes, species, serovar and saccharide type for other *w*** genes and pathway end products for pathway genes.
[c]Abbreviations: Kdo, 3-deoxy-D-manno-octulosonic acid; ECA, enterobacterial common antigen; OM, outer membrane.

Gene clusters containing ABC-type transporter genes, wzm *and* wzt *genes*

Some polysaccharide gene clusters have a pair of genes that encode proteins involved in polysaccharide export belonging to the ABC-2 subfamily of ABC-type transporters[14]. Included in this group are genes in clusters for the production of *wzy*-independent O antigens, such as *orf261* and *orf431* in *E. coli* O9 (Ref. 15) (see Fig. 2), *rfbHI* in *Vibrio cholerae* O1 (Ref. 16), *rfbAB* in *K. pneumoniae* O1 and O8 and *Serratia marcescens* O16 (Refs 17,18), and *rfbDE* in *Yersinia enterocolitica* O3 (Ref. 19). Equivalent gene pairs have also been described in several clusters responsible for capsule production and for which a common molecular origin has been

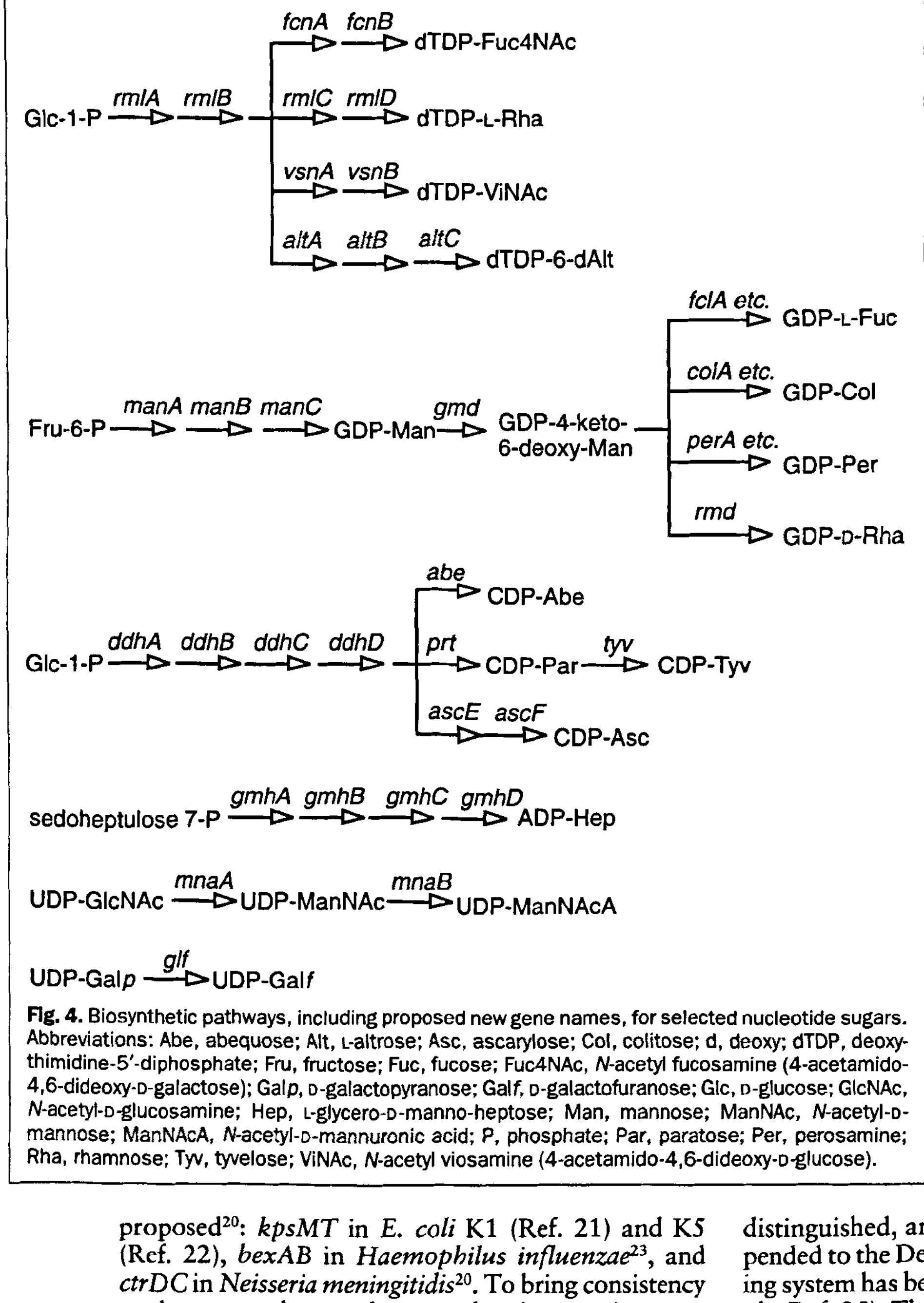

Fig. 4. Biosynthetic pathways, including proposed new gene names, for selected nucleotide sugars. Abbreviations: Abe, abequose; Alt, L-altrose; Asc, ascarylose; Col, colitose; d, deoxy; dTDP, deoxythimidine-5′-diphosphate; Fru, fructose; Fuc, fucose; Fuc4NAc, *N*-acetyl fucosamine (4-acetamido-4,6-dideoxy-D-galactose); Gal*p*, D-galactopyranose; Gal*f*, D-galactofuranose; Glc, D-glucose; GlcNAc, *N*-acetyl-D-glucosamine; Hep, L-glycero-D-manno-heptose; Man, mannose; ManNAc, *N*-acetyl-D-mannose; ManNAcA, *N*-acetyl-D-mannuronic acid; P, phosphate; Par, paratose; Per, perosamine; Rha, rhamnose; Tyv, tyvelose; ViNAc, *N*-acetyl viosamine (4-acetamido-4,6-dideoxy-D-glucose).

proposed[20]: *kpsMT* in *E. coli* K1 (Ref. 21) and K5 (Ref. 22), *bexAB* in *Haemophilus influenzae*[23], and *ctrDC* in *Neisseria meningitidis*[20]. To bring consistency to the nomenclature, these can be given *wz** names; we propose *wzm* and *wzt* (after *kpsM* and *kpsT*).

Gene clusters containing wza, wzb *and* wzc *genes*
Several EPS gene clusters have three genes in common. It has been suggested that they act in processing and/or export of EPS. As with the genes discussed above, the sequence homology is sometimes low, but consistent gene order and the pattern of homologies indicate relationships and the need for a common nomenclature.

The *E. coli* K-12 colanic acid cluster[13] and the *E. stewartii* stewartan cluster (D. Coplin, unpublished; see BPGD database, Box 2) have genes encoding related proteins that are predicted to have a signal peptidase II recognition site. We suggest that genes in this class should all be given the name *wza*. If included in the proposed scheme, the *amsH* gene of *E. amylovora*, the ORF4 gene of *K. pneumoniae* K2 *cps* region, the *epsA* gene of *Pseudomonas solanacearum* and the *exoF* gene of *Rhizobium meliloti* would also be called *wza*.

In each case, the *wza* gene is followed by two downstream genes: the first encodes a protein homologous to acid phosphatases, and the second encodes a protein with an ATP-binding sequence motif and three predicted transmembrane segments. The genes of each class show considerable similarity and we suggest that they be given the names *wzb* and *wzc*, respectively. The amino-terminal region of Wzc proteins shows some sequence similarity to Wzz (Ref. 24).

Genes of unknown function

In addition to the transferase genes discussed above, a *w**** name is appropriate for any gene that does not fit into any other category, including genes for which the function is unknown. Once characterized, the gene can be renamed in the appropriate way. Some may prefer to use ORF (open reading frame) names rather than *w**** names until function is determined.

Distinguishing homologous genes from different gene clusters or strains

Homologous genes with the same function receive the same gene name regardless of species or strain. However, in some situations genes from different sources need to be distinguished, and we suggest that a subscript be appended to the Demerec name. A form of this subscripting system has been introduced recently (see, for example, Ref. 25). The subscript could have more than one component and could include symbols denoting species (e.g. Ec for *E. coli*) or specificity of the polysaccharide (e.g. O9). Thus, in *E. coli* K-12, which has genes for an O16-specific O antigen, the gene for the first step of dTDP-rhamnose biosynthesis (see below) could be specified as *rmlA*$_{\mathrm{Ec,O16}}$. The subscripts need only be used for comparison when necessary, and then only to the minimum extent needed; for example, a discussion on O-antigen genes within *E. coli* could use *rmlA*$_{\mathrm{O16}}$. Some may find it useful to include the strain or variant name (e.g. K-12) or the type of polysaccharide that the gene helps produce (e.g. Oag for O antigen) in the subscript.

Genes involved in the production of some nucleotide sugars, for example GDP-mannose and TDP-rhamnose, are present in many gene clusters and the occurrence of dual copies of a gene within a genome is not uncommon; several instances have already been observed. For example, in *Salmonella* LT2, two genes for GDP-mannose synthesis are found in both the O-antigen and the colanic acid clusters. Until now they have been known as *rfbM* and *rfbK* in the O-antigen cluster[26], and *cpsG* and *cpsB* in the colanic acid cluster[27]. Under the BPGN scheme we have two copies of *manB* and *manC*. Other options were considered to avoid giving two genes the same name within a strain but these had their own drawbacks. Where necessary the genes can be distinguished by subscripts. When it is necessary to use a form that is easily handled by computers, for example in database entries, the information can be put in parentheses instead of in subscripts. As the species and strain are defined elsewhere in such entries it is only necessary to include information on polysaccharide type; for example, the two copies of *manB* in LT2 can be designated *manB*(CA) and *manB*(Oag).

BPGN in use

The O-antigen clusters for *E. coli* O9, *K. pneumoniae* O1 and O8, *S. marcescens* O16 and *Y. enterocolitica* O3 have several common features (Fig. 2). Each contains genes that we now call *wzm* and *wzt*, which have strong protein structural and sequence similarity and, in the case of *K. pneumoniae* O1, are known to be involved in the export of O antigen[17]. The two genes have a high level of homology and are in the same relationship to each other in each cluster; this is now obvious simply by looking at the cluster maps, but previously they were named *orf261* and *orf431* in *E. coli* O9, *rfbA* and *rfbB* in *K. pneumoniae* O1 and O8 and *S. marcescens* O16, and *rfbD* and *rfbE* in *Y. enterocolitica* O3. The same genes occur elsewhere and have been given different names and presumably would have received yet further different names in the many such gene clusters to be described in the future.

The *E. coli* O9 O unit contains only mannose, and the two mannose pathway genes are easily recognized as *manB* and *manC* (named in pathway order). Three of the remaining O9 genes encode the transferases that make the specific linkages of the O9 O unit and these have unique names. They were previously named *mtfABC* for mannose transferase, and while this may seem a better name, we would soon run out of letters when, as is likely, more than 26 mannose transferases have been discovered. The *Y. enterocolitica* O unit contains 6-deoxy-L-altrose, and once the pathway genes are identified they will be renamed *rmlAB* (encoding the same reaction as the first steps in TDP-L-rhamnose synthesis) and *altA–C* for the next three steps, which are specific to TDP-6-deoxyaltrose.

Box 1. Implementation of the BPGN scheme

We suggest that gene names should be allocated after discussion with whoever is running the database, currently Peter Reeves in Sydney. This suggestion is made, first, to avoid the possibility that two research groups find the same gene at about the same time and give it two different names or that the same name is taken by two different groups for different genes and, second, to provide an opportunity to discuss any problems in name allocation.

The scheme has been applied to O-antigen genes of *Escherichia coli*, *Salmonella*, *Klebsiella pneumoniae*, *Yersinia enterocolitica* and *Yersinia pseudotuberculosis*, LOS of *Bordetella pertussis*, lipid A and LPS core of *E. coli* and *Salmonella*, enterobacterial common antigen of *E. coli* and *Salmonella* and capsule gene clusters of *Erwinia stewartii* and *E. coli* K-12. Details, and the principles underlying name allocation, are given in a web paper released in conjunction with this paper[28].

The scheme could be applied to other gene clusters, such as those for the type 2 capsules of *E. coli* and the related clusters for capsules of *Neisseria meningitidis* and *Haemophilus influenzae*, but as yet no agreement has been reached on this. In Table 1, we have included only genes in clusters in which we are, or have been, actively involved. The BPGN proposal allows for all appropriate genes to be named, but some of those researchers approached did not wish to join in the proposal and will, for the time being, continue with the old nomenclature. Nonetheless, the 29 gene clusters listed in Table 1 include 22 for which all genes are known from sequence, and we have great confidence that the scheme will be widely adopted. Now that the BPGN is launched we hope that many others use it in naming and renaming genes. We did not find time to contact all people working in the area and concentrated on the names of genes for which the whole cluster is known. We now welcome proposals to add new gene names to the scheme, regardless of how many genes of the cluster are known, and hope that new genes will be named according to the BPGN scheme.

It is readily apparent that the O units of *Salmonella* groups B, C2 and D2, *E. coli* K-12 and *E. coli* (*Shigella*) dysenteriae all have the four-gene dTDP-rhamnose pathway genes, *rmlA,B,C,D*, which are named in pathway order but are present in each case in the order *B, D,A,C* (Fig. 3). The mannose-pathway genes *manB* and *manC* are also present in the *Salmonella* clusters in the same order as those of *E. coli* O9 (Fig. 2). The CDP-abequose pathway is present in groups B and C2 and the closely related CDP-tyvelose pathway is present in group D2. All possess genes common to both dideoxyhexose pathways (*ddhA–D*), with groups B and C2 having the gene *abe*, which is unique to the CDP-abequose pathway, and group D2 having the genes *prt* and *tyv* for synthesis of CDP-paratose and then its conversion to CDP-tyvelose, respectively. The O unit of K-12 includes Gal*f*, and the gene cluster includes the gene *glf* for conversion of UDP-galactopyranose to UDP-galactofuranose. Some of the O units of this set of O antigens contain Gal, Glc or GlcNAc, for which

Box 2. Database and more information on the web

We have set up a Bacterial Polysaccharide Gene Database (BPGD) to include the new name, old names where relevant, and function where known of all genes entered into the database, together with diagrams of the gene clusters and the oligosaccharides. The database currently has all genes referred to in this paper and we aim to add all relevant genes for which there is sufficient information. The database, written in Foxpro, is available on the world wide web (http://www.angis.su.oz.au/BacPolGenes/BPGD.html). We hope the database will facilitate the distribution of information about these genes and also ease the introduction of the new scheme. Information on the submission of suggestions for new names is available at the same site. We have also put an expanded version of this paper on the web[28] because, owing to space limitations, we have only been able to discuss some aspects of the proposal here.

Outstanding questions

- What information should a gene name convey?
- How do we know when two genes with a high degree of sequence similarity are 'the same' gene?
- How should duplicate genes within a genome be named?
- How far should we go in other areas to standardize gene names, both within and between bacterial species, especially now that more and more genes are being described in large-scale sequencing?

UDP-Gal, UDP-Glc and UDP-GlcNAc are the precursors, but as all of these are involved in other pathways in these species, the genes for their synthesis are present elsewhere on the chromosome. The transferase genes for the *Salmonella* and *E. coli* dysenteriae O antigens have been identified and named *wbaN,O,P,V,Q,R,W,Z* and *wbbP,Q,R*, respectively. Only one transferase gene (*wbbL*) has been identified for *E. coli* K-12 and the others will be among the genes named *wbbH–L*. Each transferase is unique and the details can only be found from the literature or, more easily, from the BPGD (see Box 2).

These O antigens are polymerized and exported by a different pathway from those in the first group of O antigens discussed. The gene clusters contain *wzx* genes, which encode a membrane protein thought to be the O-unit flippase, and most contain *wzy*, the O-antigen polymerase gene. The gene *wzz*, which controls the chain length of the O antigen, is present in an adjacent location in each case. It is noteworthy that the *wzy* gene of *Salmonella* groups B and D1 is anomalous because it occurs elsewhere on the chromosome.

The examples illustrated in Figs 2 and 3 include cases of pathway genes common to different gene clusters and of homologous genes of some of the saccharide-processing protein families. It is now easy to see these patterns, which before were obscured by the variety of nomenclatures used. In addition, there are several examples of *rfb* names being used for genes of different function, which will be avoided under the proposed scheme.

Benefits and drawbacks of the BPGN scheme

The BPGN scheme: (1) allows each distinctive gene to have a name that is unique but identifies it as a bacterial surface polysaccharide gene, (2) allows genes with the same function to have the same name, (3) gives pathway genes names that relate to the pathway and (4) gives distinctive names to genes for families of saccharide-processing proteins.

The BPGN will remove the confusion caused by the use of different names for genes of one function and the same name for genes of different functions. The use of specific names for pathway and saccharide-processing genes will make it easier to compare genes from different species. Increasingly, there will be a desire for, and the possibility of, computer analysis of patterns in particular classes of genes and in particular groups of organisms. This will be difficult if varying approaches to nomenclature are applied. If the proposal presented here is generally adopted, it will facilitate analysis and help those outside this exciting area to access the data without a long apprenticeship in nomenclature.

There are inevitably some drawbacks. By giving the same name to genes with the same function, all genes in a gene cluster will not have the same three-letter symbol. However, we believe that the advantages of a common name for a common function far outweigh this drawback. Perhaps the greatest problems are that the scheme entails forgetting very familiar names and that, for a while at least, the old scheme will coexist with the new. However, continuity with previous nomenclature has been retained by using the second and fourth letters to maintain a link with the old name. It is perhaps worth noting that this is not the first major change in nomenclature; LPS genes were first named *rou* genes because wild-type strains of *Salmonella*, *E. coli* and other species have a smooth colonial morphology when cultured on solid medium. Mutations in LPS genes affect the smooth phenotype and the mutants were described as rough (*rou*). Later, when it was found that they clustered in different parts of the chromosome, they were renamed *rfa* (rough A), *rfb* (rough B) and *rfc* (rough C).

Conclusions

A draft of this manuscript was discussed at the American Society for Microbiology (ASM) meeting in Las Vegas, NV, USA in April 1994. It has been through much discussion and many drafts since then. Everything included in the proposal, and many alternatives not adopted, was discussed at length. Not all decisions were easy and there are still some reservations, in particular about replacing *rfc* with *wzy*. What is clear is that change is needed because the existing terminology cannot cope. We hope that the BPGN will be widely adopted and that as it becomes familiar we will all reap the benefits of unambiguous names and ease of cross-cluster and cross-species comparisons.

Acknowledgements

The work in each lab. was supported by grants, which are acknowledged in our research publications. We thank the many people who commented on drafts and in particular the contributions of Mary Berlyn, Ken Sanderson and Kenn Rudd.

References

1 Lior, H. (1994) in *Escherichia coli in Domestic Animals and Humans* (Gyles, C.L., ed.), pp. 31–72, CAB International

2 Raetz, C.R.H. (1990) *Annu. Rev. Biochem.* 59, 129–170

3 Raetz, C.R.H. (1996) in *Escherichia and Salmonella typhimurium: Cellular and Molecular Biology* (2nd edn) (Neidhardt, F.D., ed.), pp. 1035–1063, ASM Press

4 Reeves, P.R. (1994) in *Bacterial Cell Wall* (Neuberger, A. and van Deenen, L.L.M., eds), pp. 281–314, Elsevier Science

5 Schnaitman, C.A. and Klena, J.D. (1993) *Microbiol. Rev.* 57, 655–682

6 Whitfield, C., Keenleyside, W.J. and Clarke, B.R. (1994) in *Escherichia coli in Domestic Animals and Humans* (Gyles, C.L., ed.), pp. 437–494, CAB International

7 Ørskov, F. and Ørskov, I. (1992) *Can. J. Microbiol.* 38, 699–704

8 Kenne, L. and Lindberg, B. (1983) in *The Polysaccharides* (Aspinall, G.O., ed.), pp. 287–363, Harcourt Brace Jovanovich

9 Doubet, S. and Albersheim, P. (1992) *Glycobiology* 2, 505

10 Demerec, M. *et al.* (1966) *Genetics* 54, 61–74
11 Fraser, C.M. *et al.* (1995) *Science* 270, 397–403
12 Liu, D. *et al.* (1993) *J. Bacteriol.* 175, 3408–3413
13 Stevenson, G. *et al.* (1996) *J. Bacteriol.* 178, 4885–4893
14 Reizer, J., Reizer, A. and Saier, M.H., Jr (1992) *Protein Sci.* 1, 1326–1332
15 Kido, N. *et al.* (1995) *J. Bacteriol.* 177, 2178–2187
16 Manning, P.A., Stroeher, U.H. and Morona, R. (1993) in *Vibrio cholerae and Cholera* (Wachsmuth, I.K., Blake, P. and Olsvik, O., eds), pp. 77–94, ASM Press
17 Bronner, D., Clarke, B.R. and Whitfield, C. (1994) *Mol. Microbiol.* 14, 505–519
18 Szabo, M., Bronner, D. and Whitfield, C. (1995) *J. Bacteriol.* 177, 1544–1553
19 Zhang, L. *et al.* (1993) *Mol. Microbiol.* 9, 309–321
20 Frosch, M. *et al.* (1991) *Mol. Microbiol.* 5, 1251–1263
21 Pavelka, M.S., Jr, Wright, L.F. and Silver, R.P. (1991) *J. Bacteriol.* 173, 4603–4610
22 Smith, A.N., Boulnois, G.J. and Roberts, I.S. (1990) *Mol. Microbiol.* 4, 1863–1869
23 Kroll, J.S., Hopkins, I. and Moxon, E.R. (1988) *Cell* 53, 347–356
24 Becker, A., Niehaus, K. and Puhler, A. (1995) *Mol. Microbiol.* 16, 191–203
25 Jayaratne, P. *et al.* (1994) *J. Bacteriol.* 176, 3126–3139
26 Jiang, X.M. *et al.* (1991) *Mol. Microbiol.* 5, 695–713
27 Stevenson, G. *et al.* (1991) *Mol. Gen. Genet.* 227, 173–180
28 Reeves, P.R. *et al.* (1996) *http://www.angis.su.oz.au/BacPolGenes/BPGN.html*
29 Liu, D., Lindquist, L. and Reeves, P.R. (1995) *J. Bacteriol.* 177, 4084–4088
30 Brown, P.K., Romana, L.K. and Reeves, P.R. (1992) *Mol. Microbiol.* 6, 1385–1394
31 Xiang, S.H., Hobbs, M. and Reeves, P.R. (1994) *J. Bacteriol.* 176, 4357–4365
32 Stevenson, G. *et al.* (1994) *J. Bacteriol.* 176, 4144–4156
33 Klena, J.D. and Schnaitman, C.A. (1993) *Mol. Microbiol.* 9, 393–402

Fungi, animals and us

The Mycota Vol. VI:
Human and Animal Relationships
edited by D.H. Howard
and J.D. Miller

Springer-Verlag, 1996.
DM298.00 hbk (xiv + 399 pages)
ISBN 3 540 58007 7

In the recent history of mycology, there has been an unfortunate tendency for mycologists to diverge, becoming either 'microbiologically centred' specialists in medical and veterinary mycology or 'botanically centred' specialists in saprotrophic and plant pathogenic fungi. One bonus of the molecular biological revolution of the past decade is that these two camps are becoming reunited, so it is again possible to compare activities of fungi in different environments. Thus, mycologists of all persuasions will find material of interest in this volume. It is the sixth in the eight-volume series entitled *The Mycota, a Comprehensive Treatise on Fungi and Experimental Systems for Basic and Applied Research.* The most-notable comparable multivolume treatment was *The Fungi* by Ainsworth and Sussman (1965–1973), which documented the academic research at the time, with detailed treatments of physiology, cytology, life cycles and taxonomy. In recent decades, the tenor of research has changed, lead by the revolution of molecular biology, so that we now have the complete genome sequence of *Saccharomyces cerevisiae*, and genes from other fungi are being sequenced at an increasingly rapid pace. This enables us to look at the growth and activities of microorganisms in new lights and, while the approach of *The Mycota Vol. VI* is to look at the biochemical basis of interactions between fungi and their animal hosts, few chapters in this volume are untouched by the molecular aspects of these relationships.

The bulk of the book reviews medical and veterinary mycology, with eight chapters on fungal pathogenesis and five on allergenic and toxigenic effects. The first chapter sets the scene by reviewing the wide range of fungal factors associated with pathogenesis and is followed by a detailed account of the involvement of enzymes, especially proteases, in this process. Chapters 2 and 3 consider cell-mediated and humoral immunity, respectively: fields that have received much impetus because of the prevalence of fungal infections in AIDS patients. The major human fungal pathogens, particularly those associated with HIV infection, are then dealt with in detail. In addition to their pathogenesis by growth in human and animal tissue, fungi have major detrimental effects on our lives and those of our farm animals and pets by more insidious means: by the inhalation of their spores or mycelial fragments, giving rise to 'organic-dust toxic syndrome' and allergic responses, and by ingestion of mycotoxins. The characterization of these effects is notoriously difficult in the field and they are probably often unrecognized, which is a good reason for devoting nearly a quarter of this volume to them.

In evolutionary terms, fungi and invertebrates have coexisted for very much longer than fungi and vertebrates. Thus, it is not surprising to find a much more diverse range of associations between the former groups than between the latter, which are almost totally confined to a relatively small number of pathogenic fungi. One exception, where specific coevolution has occurred, is the occurrence of the anaerobic chytrids that inhabit the guts of all ruminants and those of many other herbivorous mammals. These remarkable microorganisms are discussed within a comprehensive, well-illustrated chapter (Chapter 14). Much older associations are shown by the many types of interactions between fungi and invertebrates. The most studied are the interactions with insects, and three chapters consider these. Chapter 16 describes the strange Trichomycetes, probably a polyphyletic grouping, which are very highly adapted to their obligate life in arthropod guts, while Chapters 17 and 18 deal with pathogenic, mutualistic and