BIOCHEMICAL JOURNAL LETTERS

Nomenclature for sugar-binding subsites in glycosyl hydrolases

The huge structural diversity of polysaccharides leads to their central roles in food storage and utilization, structure, cell–cell signalling, cell-wall expansion and turnover and viral invasion. Glycosyl hydrolases, enzymes hydrolysing the glycosidic bond in di-, oligo- and poly-saccharides, are found in all living organisms. The first X-ray structural determination of an enzyme was of a glycosyl hydrolase: hen egg-white lysozyme (HEWL) [1]. Since then, over 57 sequence-based families of glycosyl hydrolases have been identified [2], and three-dimensional structures are known for representatives of over 22 of these [3]. This rapid growth of known three-dimensional structures of glycosyl hydrolases has been accompanied by a diverse and disparate array of nomenclature for the labelling of their sugar-binding subsites.

In a number of depolymerizing enzymes, catalytic activity is affected by substrate-binding sites distant from the bond actually undergoing hydrolysis. Such a subsite system is not only encountered in glycosyl hydrolases, but also in proteinases and nucleases. The number of subsites, the energy of interaction of each subsite and the hydrolytic rate coefficients may be determined experimentally [4]. There is clearly a need for an appropriate and consistent nomenclature for the labelling of the subsites in glycosyl hydrolases. Sadly, the current literature is beset with problems regarding the naming of the enzyme subsites which bind saccharides. Whilst most enzymologists have chosen to use one system, the crystallographic community boasts almost as many nomenclatures as there are published papers. Comparisons between various complexes of the same enzyme are difficult, and between different enzymes, almost impossible. The fundamental basis for a consensus nomenclature must be that it allows comparison both between different enzymes with different numbers of subsites and between several complexes of the same enzyme. Two criteria are essential: it must indicate the position of the subsite relative to the point of cleavage and must not change the subsite labelling when new complexes, with extra sugar units at either the reducing or the non-reducing end, become known.

We propose that the structural-biology community adopts the $-n$ to $+n$ subsite nomenclature widely used by molecular enzymologists. Subsites are labelled from $-n$ to $+n$ (where n is an integer). $-n$ represents the non-reducing end and $+n$ the reducing end, with cleavage taking place between the -1 and $+1$ subsites. Before detailing this nomenclature, we place the various labelling schemes for polysaccharide-degrading enzymes into a historical perspective and address their respective strengths and weaknesses.

Enzymological mapping of glycosyl hydrolases

Subsite mapping of glycosyl hydrolases began in the late 1960s with seminal studies on amylolytic enzymes. Enzyme subsites were labelled i, $i+1$, $i+2$ to $i+n$ etc., with the numbers increasing positively towards the reducing end of the substrate. The reason was that, in sugar chemistry, oligomers are drawn by convention with the non-reducing end on the left-hand side and the reducing end on the right. Early work on an exo-amylolytic enzyme [5] labelled the subsites 1, 2, 3, 4 and 5 (from non-reducing to reducing end), with cleavage taking place between subsites 1 and 2 (Figure 1a). This system was appropriate for exoamylases, which cleave the non-reducing terminal sugar, because hydrolysis always occurs between the same two subsites. As work extended to endo-acting enzymes, however, this nomenclature led to increasing confusion, since one enzyme could cleave between subsites 5 and 6 and another between 6 and 7. Additionally, the discovery of subsites beyond the 1 subsite forced some authors

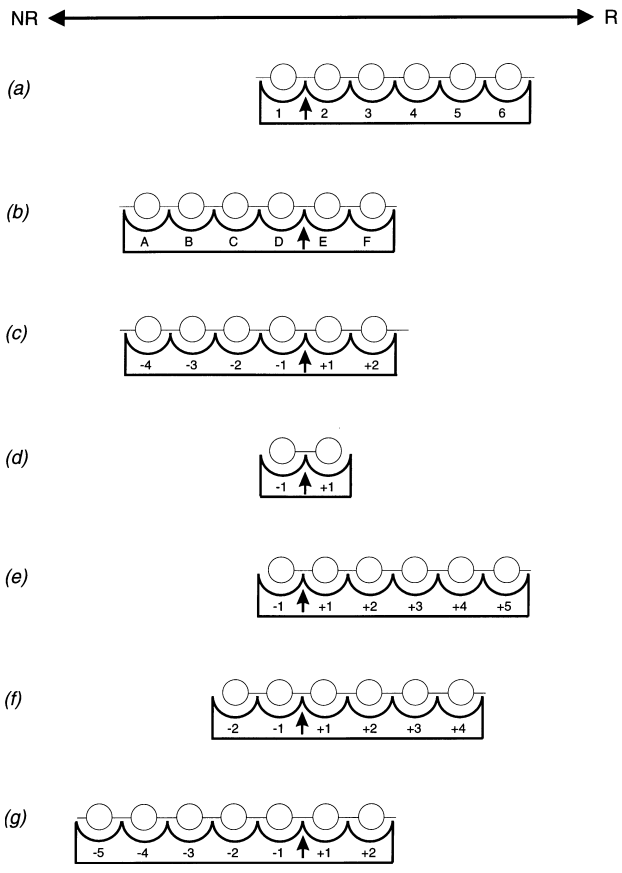

Figure 1 Schematic drawing of the sugar-binding subsites in several glycosyl hydrolases

By convention, the non-reducing end of the substrates is drawn on the left and the reducing end on the right. The point of cleavage is indicated by an arrow. (**a**) early subsite labelling by numbers, as applied to glucoamylase; (**b**) subsite labelling by letters, as applied to HEWL; (**c**) the subsites of HEWL labelled with the proposed $-n$ to $+n$ scheme; (**d**) the $-n$, $+n$ system applied to non-specific monoglycosidases and disaccharidases; (**e**) the $-n$, $+n$ system applied to enzymes cleaving a monosaccharide from the non-reducing end of the substrate such as glucoamylase; (**f**) the $-n$, $+n$ system applied to enzymes cleaving disaccharide units from the non-reducing end of the substrate such as β-amylase; (**g**) the $-n$, $+n$ system applied to enzymes cleaving disaccharide units from the reducing end of the substrate as proposed for *T. reesei* cellobiohydrolase I.

to include negative and obscurely labelled subsites [6]. Similar problems also arise with all alphabetically based systems.

In 1978 Suganuma and co-workers, whilst still using the 1, 2, 3, … *n* nomenclature, simultaneously introduced the concept of labelling of subsites relative to the point of enzymic cleavage [7]. Cleavage was designated as taking place between the *r* and *r*+1 sites, with the subsites after the point of cleavage towards the reducing end having positive numbers (*r*+1, *r*+2 … *r*+*n*), with more negative values towards the non-reducing end (*r*−1, *r*−2 etc.). Since then, most subsite mapping work has adopted a refined form of this nomenclature, changed to improve consistency and to allow the direct comparison of results on different enzymes. Subsites are labelled such that −*n* represents the non-reducing end, *n* the reducing end, and cleavage occurs between the −1 and +1 subsites. Thus enzyme subsites towards the reducing end of the substrate are labelled +1, +2, +3 to +*n* and those towards the non-reducing end, away from the point of cleavage, −1, −2, −3 to −*n*. It is unclear when this subtle, but extremely powerful change in the nomenclature occurred, but it is now a common means of description prevalent in the literature (see, for example, [8–10]).

Crystallographic work on lysozyme: the alphabetical system

The cardinal paradigm for the labelling of glycosyl hydrolase subsites by protein crystallographers is the alphabetical nomenclature for HEWL introduced by Phillips and co-workers [11]. Analysis of the structure of HEWL, together with early studies on complexes, identified six subsites for saccharide binding, labelled A–F, with A at the non-reducing end and F at the reducing end of the sugar substrate. The cleavage point was between the D and E subsites (Figure 1b). Various versions of this alphabetical system are used for most published structures of glycosyl hydrolase complexes.

Whilst this scheme is suitable for describing the subsite structure of a single enzyme, it is inappropriate for the comparison of different enzymes and their complexes. Firstly, a compound may bind to different subsites in related enzymes, leading to apparently similar, but in reality different, nomenclatures. An example of this is the binding of the inhibitor acarbose to the amylolytic family 13 enzymes. The A and B sugars from one study [12] occupy the same subsites as the B and C residues of another [13]. Also, after the first complex of an enzyme has been solved and the subsites labelled alphabetically, later work frequently reveals subsites beyond the A subsite. As in the 1–*n* numerical systems, this forces the use of counter-intuitive names for the extra subsites and confusion in the published literature. Further ambiguity also arises when some groups use reverse alphabetical systems starting at the reducing end, often in contradiction with previously published work on the same system.

Finally, as a direct consequence of the first two problems, descriptions and comparisons of the enzymic reaction mechanism become tiresome. The subsite where catalysis takes place receives many different names. For example, in an alphabetical system growing from non-reducing to reducing end, the catalytically equivalent subsite is labelled A in glucoamylase from *Aspergillus awamori*, B in cellobiohydrolase-II from *Trichoderma reesei*, C in endoglucanase CelA from *Clostridium thermocellum*, D in HEWL and E in cellobiohydrolase-I from *T. reesei*.

International Union of Pure and Applied Chemistry–International Union of Biochemistry (IUPAC–IUB) nomenclature for polysaccharide chains

There is no reported structural work using the IUPAC–IUB guidelines for the nomenclature of polysaccharide chains. In the IUPAC-IUB system, polysaccharide chains are numbered from the reducing end to the non-reducing end [14] The reason for this is so that '… the gain or loss of a residue at the non-reducing end … does not change the numbering of every unit in the chain …'. In an enzyme-catalysed reaction, it is the location of the subsite relative to the point of enzymic cleavage that is important, not the position relative to either end of the polysaccharide chain. Indeed in describing a polysaccharidase, description of the cleavage point relative to a position that could be randomly tens or hundreds of units away from the reducing end of the chain is impossible. A second pre-requisite for a subsite nomenclature is that the addition or loss of a glycosyl unit at either end of the chain does not change the labelling of every subsite. A simple transfer of the IUPAC–IUB polysaccharide chain numbering on to the enzyme subsites would result in a completely inconsistent subsite nomenclature in cases where alternate complexes display different binding modes. A further complication arises when two molecules are simultaneously bound in different parts of the active site [15], since there are two reducing ends which would be both labelled 1. The IUPAC–IUB nomenclature was designed for the labelling of polysaccharide chains and is not appropriate for description of enzyme active sites.

Proteinases

The first enzymes to acquire a consistent subsite nomenclature, used by both X-ray crystallographers and enzymologists, were not glycosyl hydrolases, but serine proteinases. Ironically, it was the revelation of the active-site cleft and sugar-binding subsites in HEWL that led Schechter and Berger [16] to propose the subsite nomenclature for proteinases. In this system, the proteinase subsites for amino-acid binding are defined such that cleavage takes place between the S_1 and S'_1 subsites, with primed subsites S'_1, $S'_2 \ldots S'_n$ indicating subsites towards the C-terminal end of the substrate and the unprimed subsites S_1, $S_2 \ldots S_n$ representing those subsites towards the N-terminal end of the substrate.

The simple proteinase subsite nomenclature with primed subsites 'after' the point of cleavage also fulfils the essential criteria and, with hindsight, it is somewhat unfortunate that this nomenclature was not adopted by the glycosyl hydrolase community at an early stage. To date only a handful of structural papers have used this nomenclature [17]. More recently, one group has adopted a nomenclature based on that used for the proteinases, but replaced the prime (′) sign with a minus (−) [18]. Whilst there is nothing formally incorrect with this system, it has resulted in the extreme confusion of a nomenclature that uses −*n* to +*n*, with cleavage between −1 and +1, but with the opposite value of positive or negative to that used by other published work with a seemingly identical nomenclature.

Proposed nomenclature

We propose that the structural-biology community use the system in which subsites are labelled from −*n* to +*n*, with −*n* at the non-reducing end and +*n* the reducing end. Cleavage occurs between the −1 and +1 subsites. The wide applicability of this nomenclature is best illustrated with a series of examples which show that all enzyme classes are now comparable.

1. Endo-polysaccharidases: the A–F subsites of HEWL become subsites −4 to +2 (Figure 1c). The −*n*, +*n* nomenclature would likewise be applicable to all endo-polysaccharidases such as endoglucanases and chitinases, and has already been applied in several structural studies [9,15].

2. Glycosidases: those glycosidases which are specific for a certain sugar, but less specific for the aglycone, such as β-galactosidase, have only two subsites, −1 and +1, but with

little specificity for $+1$ (Figure 1d); similarly, disaccharidases, such as chitobiase, are -1, $+1$ enzymes but with greater specificity for a given saccharide in the $+1$ subsite (Figure 1d).

3. Exo-polysaccharidases: those which cleave off monosaccharides from the non-reducing end of a polymeric substrate, such as glucoamylase, are -1, $+n$ enzymes with $n > 1$ (Figure 1e); similarly an exo-polysaccharidase which liberates disaccharides from the non-reducing end, such as β-amylase, has subsites -2 to $+n$ with $n > 2$ (Figure 1f); and a cellobiohydrolase which is proposed to liberate cellobiose from the reducing end of the polymer, such as *T. reesei* CBH-I, has subsites labelled $-n$ to $+2$ (Figure 1g).

We believe that this is a self-consistent description of the sugar-binding subsites for this important set of saccharide-metabolizing enzymes. Its use should allow straightforward comparison of the active sites of all such enzymes with respect to the point of cleavage. We hope that its use will be rapidly accepted by the community.

Gideon J. DAVIES*‡, Keith S. WILSON* and Bernard HENRISSAT†

*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), C.N.R.S., BP 53, F-38041 Grenoble Cédex 9, France

‡ To whom correspondence should be addressed.

1 Blake, C. C. F., Koenig, D. F., Mair, G. A., North, A. C. T., Phillips, D. C. and Sarma, V. R. (1965) Nature (London) **206**, 757–763
2 Henrissat, B. and Bairoch, A. (1996) Biochem. J **316**, 695–696
3 Davies, G. and Henrissat, B. (1995) Structure **3**, 853–859
4 Allen, J. D. (1980) Methods Enzymol. **64**, 248–277
5 Hiromi, K. (1970) Biochem. Biophys. Res. Commun. **40**, 1–6
6 Thoma, J. A. and Allen, J. D. (1976) Carbohydr. Res. **48**, 105–124
7 Suganuma, T., Matsuno, R., Ohnishi, M. and Hiromi, K. (1978) J. Biochem. (Tokyo) **84**, 293–316
8 Biely, P., Vrsanska, M. and Claeyssens, M. (1991) Eur. J. Biochem **200**, 157–163
9 Hrmova, M., Garrett, T. P. J. and Fincher, G. B. (1995) J. Biol. Chem. **270**, 14556–14563
10 Planas, A. and Malet, C. (1995) in Carbohydrate Engineering (Petersen, S. B., Svennson, B. and Pedersen, S., eds.), pp. 85–95, Elsevier, Amsterdam
11 Blake, C. C. F., Johnson, L. N., Mair, G. A., North, A. C. T., Phillips, D. C. and Sarma, V. R. (1967) Proc. R. Soc. London B **167**, 378–388
12 Qian, M., Haser, R., Buisson, G., Duée, E. and Payan, F. (1994) Biochemistry **33**, 6284–6294
13 Strokopytov, B., Penninga, D., Roseboom, H. J., Kalk, K. H., Dijkhuizen, L. and Dijkstra, B. W. (1995) Biochemistry **34**, 2234–2240
14 IUPAC–IUB (1983) Eur. J. Biochem **131**, 5–7
15 Davies, G. J., Tolley, S. P., Henrissat, B., Hjort, C. and Schülein, M. (1995) Biochemistry **34**, 16210–16220
16 Schechter, I. and Berger, A. (1967) Biochem. Biophys. Res. Commun. **27**, 157–162
17 Klein, C., Hollender, J., Bender, H. and Schulz, G. E. (1992) Biochemistry **31**, 8740–8746
18 Strokoptyov, B., Knegtel, R. M. A., Penninga, D., Rozeboom, H. J., Kalk, K. H., Dijkhuizen, L. and Dijkstra, B. W. (1996) Biochemistry **35**, 4241–4249

Received 5 September 1996