Glycoconjugate Journal (1993) **10**:419–421

GUEST EDITORIAL

The nomenclature of glycosaminoglycans and proteoglycans

JOHN E. SCOTT
Department of Chemical Morphology, Manchester University, Manchester M13 9PL, UK

Introduction

Terminologies (including abbreviations and acronyms) in these fields are *ad hoc*. Some recall when 'mucopolysaccharide' and 'mucoprotein?' were the best available terms, before the blessed Balazs and Jeanloz introduced 'glycosaminoglycan' (GAG) and 'proteoglycan' (PG), together with a clean-up of endings (kerat*an*, dermat*an*, hepar*an*) which gave a chemical gloss to our basic ignorance. Things have moved on since then, but the terminology has not.

In the past few years names such as decorin, lumican, aggrecan, syndecan, etc. have been given to molecules whose chemistry was known in detail. These names lack chemical information and internal consistency. The need for new tools to solve problems includes ways of expressing results so that we understand, even if with some effort, what is done by others. Imagine working with 'common salt' and 'oil of vitriol', without the symbols (or even names) for sodium, sulphur, etc.

The 8th Harden Discussion held in Manchester (March 28–29, 1993), on 'Dermatan sulphate proteoglycans' was an outstanding opportunity to seek bases for economic and unambiguous communication. Registrants were asked for proposals, but since none was received (not unusual), I circulated my own draft six weeks before the meeting, and this produced considerable feedback, which was circulated as two further appendices. A lively two hour discussion of these documents at the meeting was recorded. The outcome is in the book *Dermatan Sulphate Proteoglycans, Chemistry, Biology, Chemical Pathology* (J. E. Scott, ed.) published in November 1993 [1].

Glycosaminoglycan terminology

The pioneering work of Karl Meyer identified polymers with a similar mix of repeating units, based on fractionation and analytical data. His terms (e.g. chondroitin sulphates A, B or C) defined major tissue components. They were of great value, but difficulties now grow with increase of knowledge:

(i) hybrid polymers from many tissues do not fit into this system,
(ii) domain structures [2] present in many (maybe most) tissue GAG are not recognized,
(iii) the spectra of modifications due to sulphation and epimerization provide no basis on which to distinguish between, for example, chondroitin and dermatan sulphates (CS and DS). If 10% iduronate (IdoUA) qualifies CS to be called DS, are CS containing 9% iduronate and DS containing 11% iduronate different species?
(iv) molecular recognition is dependent on saccharide sequences and domain structures.

My proposal was that terminology be based on disaccharide units, which are readily accessible to quantitative analysis, via enzymic digestion. These units are of unambiguous composition and can be represented by one-letter codes.

Polymer abbreviations should be two-letter codes, e.g. CS, DS, HS (heparan sulphate), KS (keratan sulphate). If there is no sulphation, Ch, De, He and Ke are used. They are defined in terms of disaccharide units, so Ch (chondroitin) is a polymer of repeating disaccharides of –4GlcUAβ1–3GalNAcβ1–.

DS (currently an abbreviation for dermatan sulphate) contains uronic acid C5 epimers of CS disaccharides. Probably all 'DS' contains CS units. To avoid confusion about *definitions* of dermatan sulphate (implying epimerization of all glucuronic acid residues to iduronic acid residues) and actual isolates from tissues, a new term is proposed, 'dermochondan sulphate' [3], indicating that 'DS' preparations are co-polymeric. The abbreviation 'DS' is used for these polymers, currently called dermatan sulphate.

KS consists of disaccharides of –3Galβ1–4GlcNAcβ1– units, sulphated to various extents and in different positions. It belongs to the same polymer group as CS [4].

HS is the sulphated polymer of heparan (He) containing repeating disaccharides of –4GlcUAβ1–4GlcNAcα1–

or –4IdoUAα1–4GlcNAcα1–. It is therefore analogous to demochondan sulphate by containing two uronic acid epimers. There are no names analogous to dermatan or chondroitin in this GAG family.

One-letter codes for disaccharides

Four positions on the repeating Ch disaccharide can be sulphated, giving $2^4 = 16$ possible disaccharides. At least nine have been characterized. Existing alphanumeric terminology and abbreviations will be stretched if more than three different disaccharides are found in long sequences. Work on DS implies that sequences will be written out like those in polypeptides ([1], pp. 11–41). Molecular recognition phenomena require that the participants be described succinctly. We need a simple code system.

I proposed that disaccharides be labelled by one-letter codes, according to the position(s) of their sulphate group(s). By using disaccharides as units, details of glycosidic bonds are subsumed into a single letter. Importantly, GAG analyses depend on enzymes (chondroitinases, etc.) which produce these disaccharides.

Allotting one-letter codes to relevant units

In contrast to the well-established one-letter amino acid code, non-arbitrary allotment of letters to disaccharides is possible. The code for a given disaccharide can be derived, without committing it to memory.

The disaccharides are listed in alphabetical order, following the sequence, (i) low to high sulfation, (ii) low ring numbers to high ring numbers.

(a) The letter 'A' is used for unsulfated disaccharides [5].
(b) The four monosulphates then use the letters B–E. Since sulphation is predominantly in the *N*-acetylglucosamine residue, lettering starts on this ring; B = 4 sulphation, C = 6 sulphation, D = sulphation at C2 in the glucuronic acid ring, E = GlcUA C3 sulphation.
(c) The six possible disulphates use F–K, F = sulphation at C4 and C6 of the galactose ring, G = sulphation at GalNAc C4 and GlcUA C2 (taking the low numbers first).
(d) The same rules apply to the four trisulphates and to the tetrasulphate (P) (see [1] pp. 1–11 for further details).

Many CS preparations are hybrids, conveniently expressed for example, CS(A, F, C), where the left-to-right order indicates decreasing amounts of the relevant disaccharides. A prevalence of letters in the later alphabet indicates higher degrees of sulphation.

Quantitative information can be expressed e.g. CS(A50, C30, G20), where the figure denotes percentage of A etc. The system can show exact molar composition, e.g. CS(A_{20}, C_{12}, F_4), best restricted to samples containing only one species of CS molecule. In this example all units are CS units, i.e. GlcUA-containing.

One letter code sequences

The A, B, C, etc. codes can be used to show sequences of disaccharides in polymers or oligomers. For example, AAAEEE is a dodecasaccharide comprising A and E disaccharides in the sequence shown, with the non-reducing end to the left. To distinguish between GlcUA- and IdoUA-containing units, for example in a DS sequence; a symbol rather than a letter (which could be confused with other letters) is suggested [6]. The combination should be machine-readable [7]. The +(=GlcUA) and −(=IdoUA) signs (A^+, C^-, etc.), are suggested. If the GlcUA/IdoUA status of the sugars is not known, no qualifying sign is needed (e.g. in the case of disaccharides from an ABC lyase digestion).

Extension of one-letter codes to KS and HS

All GAG polymers consisting of repeating disaccharide units can be similarly systematized (see [1], pp. 1–11 for discussion). If sequences from different GAGs are used simultaneously, they could be distinguished for example as CS(AAAEEE) and KS(AAAEEFA).

Terminology of proteoglycans

The current unsatisfactory situation has been briefly described [8]. A rational system should convey information about the protein *and* the glycan parts. Single names purporting to describe both are certain to confuse, since one part is a gene product and the other is post-translational. They do not necessarily occur together.

The abbreviation PG is in wide use. It is consistent to use proteochondroitin sulphate (PCS), proteokeratan sulphate (PKS), and now proteodermochondan sulphate (PDS) as abbreviations for PGs with CS, KS or DS chains, respectively. If more than one type of GAG chain is attached to the protein, it is expressed, for example, as PCS, KS or PCS, HS. The dominant GAG is stated first. This convention can include quantitative or semi-quantitative information about the GAG, e.g. PCS(A, F) (see above). It accommodates data on numbers of GAG chains attached to the protein e.g. $P(DS)_{7-10}$; $P(CS)_{70-100}$; $(KS)_{10-20}$.

Protein cores may be viewed as gene products, as amino acid sequences, as functioning units, or as characteristic shapes (sizes).

Proposals

(i) Names in current use, e.g. decorin, should be used to describe only the gene product.
(ii) To emphasize their connection with the gene, rather than with the glycan, the ending 'on' (as in exon, intron, codon) should replace '-an' etc. [9]. Thus; decoron, lumicon.

(iii) A PG is indicated by adding appropriate GAG abbreviation(s), e.g. decoron DS, lumicon KS, aggrecon CS, KS.

In summary

1. The proposed terms and codes provide concise, quantitative information on the polymer backbone, state of oxidation, patterns of sulphation and epimerization, and proportions of monomeric units. Oversulphated domains and special units are easily recognized.

2. They are intermediate between general statements (e.g. 'chondroitin sulphate') and detailed primary structures.

3. They bring terminology into line with current analytical techniques, which depend largely on enzymes developed by S. Suzuki and co-workers. Current enzyme nomenclature (chondroitinase ABC, AC) can still be used.

4. They are not primarily intended to be spoken, although some codes are easily articulated.

5. They have capacity and flexibility to accommodate future developments.

Invitation

Contributions to a continuing discussion are warmly invited. Send letters to J. E. Scott, at the above address.

Acknowledgement

My thanks are due to Dr Alan Chester for help during the preparation of this document.

References

1. Scott JE (1993) (ed) *Dermatan sulphate proteoglycans. Chemistry, Biology, Chemical Pathology*. London: Portland Press.
2. Oeben M, Keller R, Stuhlsatz HW, Greiling H (1987) *Biochem J* **248**:85–93.
3. Silbert JE. Suggestion.
4. Scott JE (1991) *Biochem J* **275**:267–68.
5. Suzuki S. Suggestion.
6. Yoshida K. Suggestion.
7. Scott PG. Suggestion.
8. Scott JE, Greiling H, Linker, A (1992) *Trends Biochem Sci* **17**:176–77.
9. Smith RKW. Suggestion.