Pure & Appl. Chem., Vol. 69, No. 9, pp. 1867-1873, 1997.
Printed in Great Britain.

Information capacity of the carbohydrate code

Roger A. Laine

Departments of Biochemistry and Chemistry, Louisiana State University and A&M College and The Louisiana State University Agricultural Center, Baton Rouge, Louisiana USA 70803

Abstract: Capacity for information in biological molecules is traditionally thought to reside in the primary sequence of proteins and RNA, recorded in the DNA. With the exception of some RNA molecules, proteins, if not structural, carry their information in binding sites for substrates of reactions, or in binding sites for control molecules. Some proteins bind to complex carbohydrates in a carbohydrate-specific fashion, including enzymes, lectins and antibodies. These carbohydrates, assembled by sequential glycosyl transferases, also carry biological information, the other side of which is a binding protein that recognizes a specific sugar monosaccharides, sequence, anomerity, linkage, ring size, branching and substitution. It is the latter 7 parameters, however, that give carbohydrates a very large potential for information-carrying capacity in a short sequence. An exponentially growing body of knowledge exists in this aspect of carbohydrate function.

INTRODUCTION

Carbohydrates contain an evolutionary potential of information content several orders of magnitude higher in a short sequence than any other biological oligomer [1, 2] This is due to monomers capable of more than one linkage position, anomerity and branching. This high potential for information capacity exists in biological recognition systems comprised of complex carbohydrate ligands on the one hand which are recognized for targeted activities, on the other hand, by hapten specific protein receptors, such as lectins or antibodies. Evidence is accumulating that carbohydrate-carbohydrate interactions play roles in some recognition systems [3-13]. Certainly this is true in self-association of polymer chains such as cellulose and chitin.

In biology, the oligosaccharide subunit size recognized by a binding protein is commonly 6 sugars or fewer, usually 1-4. The protein cognate receptor can be one of the following:

- soluble lectin (non-immune carbohydrate-binding protein)
- antibody
- hemagglutinin
- cell surface lectin
- enzyme

* (or another carbohydrate)

There are 7 elements to the haptenic character of carbohydrates as follows:

- monosaccharide identity: epimer
- anomeric configuration: a, ß
- linear sequence
- ring size (furanose or pyranose)
- position of linkage on ring
- branching pattern
- substitutions
    - phosphate, phosphonate
    - sulfate, sulfonate
    - alkyl, acyl, acetal, others

On the protein side of the cognate pair, the carbohydrate-binding protein, there are also several elements in the properties of the binding site, namely:

- number of monosaccharide subunits fitted in the binding site;
- preferred 3 dimensional form of the oligosaccharide;
- details of each saccharide subunit recognized (linkage, anomeric, ring size);
- binding site acceptance to alternate ligands (specificity);
- valency of receptor.

The importance in biology of the lectin/carbohydrate recognition pair is becoming increasingly evident by the >8000 articles published in the past 16 years which use the word "lectin" in the title or abstract. Figure 1 shows the trend of research in this area by 5 year increments as represented by the number of publications containing the word "lectin" in the title or abstract.

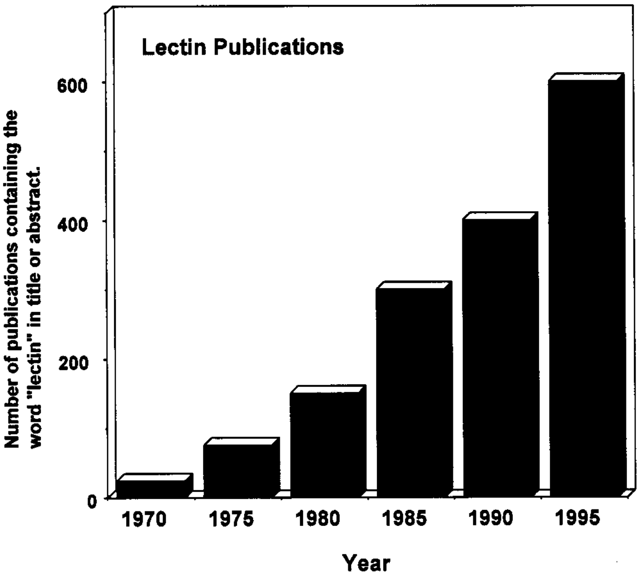

Figure 1

A typical calculation

Consider that a trisaccharide could be made up of any of the 20 most commonly found sugars such as glucose, mannose, galactose, fructose, N-acetylglucosamine, N-acetylgalactosamine, fucose, arabinose, xylose, ribose, glucuronic acid, galacturonic acid, mannuronic acid, iduronic acid, sialic acid, KDO, KDN and others.

The number of possible unsubstituted *trisaccharides* would be as follows:

[(permutations of sequence) x anomeric x ring size x linkages]

or

[20^3 x 2^3 x 2^3 x 12]

(average for potential linkage position isomers is around 12 for 3 sugars) or

>6,000,000 linear (and 3,000,000 branched) structures as shown below)

This gives a total of 9 x 10^6 potential trisaccharide structures using a library of only 3 sugars*vs* only 8000 total for a tripeptide made from 20 amino acids (20^3). This is 3 orders of magnitude difference.

Most commonly, biologically recognized complex carbohydrate sequences as either repeat units in polysaccharides or as oligosaccharides are composed of sets of 4 or fewer epimers of common hexoses or pentoses, 0-2 different amino sugars, 0-1 methyl pentose and 0-1 sialic acid, ketodeoxynononic acid (KDN), ketodeoxyoctonic acid (KDO) or uronic acid. There also occur in microbes and plants many deoxy and dideoxy sugars, glycosaminuronic acids and branched sugars like apiose.

In many systems oligosaccharide sequences are substituted with functional groups such as sulfate, methyl or acetate, usually in a structural motif of fewer than 7 sugars.

From a set of 3 hexoses, use the formula:

$$\textbf{Structures} = \mathbf{E^{n} \text{ x } 2^{n}{}_{r} \text{ x } 2^{n}{}_{a} \text{ x } 4^{n-1}}. \qquad (1)$$

$\mathbf{E^n}$ represents the permutations from order of sequence including repetitions of the same sugar $3^3 = 27$.

Where **E** is the library of sugars (3 in this special case), and **n** is the oligomer size (also 3 in this case).

The total is multiplied by another term for ring size, $\mathbf{2^n{}_r}$ or $2^3 = 8$ since most sugars can occur in nature as either pyranose or furanose forms. The total is again multiplied by $\mathbf{2^n{}_a}$, a term for anomeric configuration: $2^3 = 8$. The linkage position term is relevant for 2 of the 3 sugars, hence $\mathbf{4^{n-1}}$ where 4 potential hydroxyls are available for linkage to the previous sugar and gives a number $4^2 = 16$. For pentoses, the term would have been $3^2 = 9$, therefore we are assuming an average of 12.

With furanose rings in a trisaccharide of sequence ABC, sugar 'A' could have been connected through the 5 position of sugar 'B', for example. This factor is covered, however, by the ring size term $\mathbf{2^n{}_r}$ keeping the total possibilities of linkage positions for hexoses at 16. Thus, the correct number for permutations of linear trisaccharides made up from a set of 3 hexoses is

27 x 8 x 8 x 16 = 27,648.

Branching

```
A(1->6)                 B(1->6)
     \                       \
      C(1->R)*    or          C(1->R)*
     /                       /
B(1->3)                 A(1->3)
```

*R = reducing end attachment site (aglycon)

In branched trisaccharides sugars A and B are both glycosides to sugar C, branched in 6 possible combinations as 2,3; 2,4; 2,6; 3,4; 3,6 or 4,6. If sugar C were in the furanose form, additional isomers include 2,3; 2,5; 2,6; 3,5; 3,6; or 5,6 for a total of 12 different branched structures. However, the ring size term $\mathbf{2^n{}_r}$, accounts for the additional 6 structures if C were furanose. Since each branch can occur in two different ways, such as A1->6;B1->3 or B1->6;A1->3 there are 12 different ways to branch these three sugars. The permutation term, $\mathbf{E^n}$, however, accounts for the A6;B3 and B6;A3 branching duplex.

Therefore, the unique set of branched trisaccharides from a set of 3 hexoses are **27 x 8 x 8 x 6 = 10,368.**

The total structures from a trisaccharide comprised of 3 hexoses, choosing among a set of only 3 different hexoses, is 27,648 (linear forms) plus 10,368 (branched forms) = 38,016 (from 3 amino acids, only 27).

The formula for isomers of a trisaccharide having a reducing end is thus:

$$\mathbf{E^{n} \text{ x } 2^{n}{}_{r} \text{ x } 2^{n}{}_{a} \text{ x } 4^{n-1} \text{ (linear forms)} = 27{,}648} \qquad (2)$$

$$+$$

$$\mathbf{E^{n} \text{ x } 2^{n}{}_{r} \text{ x } 2^{n}{}_{a} \text{ x } 6^{n-2} \text{ (branched forms)} = 10{,}368} \qquad (3)$$

$$\mathbf{\text{Total} = 38{,}016}$$

The 3 dimensional presentation of carbohydrate structures to epitope-specific recognizing proteins comprises a "high level language" biochemical code. In this view, DNA can be considered as "machine language", coding for the lectins or antibodies on one hand and on the other hand the sets of glycosyl transferases that assemble the sugars.

Antibodies are a prime example of binding proteins, being exquisitely sensitive to all 7 of the carbohydrate structural elements. The isomer permutations of small (M_r <1500) carbohydrates are 3-4 orders of magnitude larger than peptides at the trisaccharide level, and 7 orders of magnitude larger at the hexamer level, the large increase due to saccharide branching.

NON-reducing oligosaccharides: Trisaccharides can also be configured as the trehalose-type disaccharide aldose-1->1-aldose or the sucrose-type non-reducing aldose-1->2 ketose diacetal linkage structure, producing a larger isomer set. Longer oligosaccharides can self-bond end-to-end forming cyclodextrins, adding more to the potential isomers. There are still questions whether many of these types of molecules possess biological activity other than energy and carbon storage. Certainly the acylated trehaloses of mycobacteria cell walls, called "cord factor" are potent immunogens.

Substitutions

Often, carbohydrates are substituted with functional groups. Returning to the example of trisaccharides made up from a library of 3 hexoses, each member of 38,000 isomeric structures could be substituted, for example, by one sulfate in any of 10 free hydroxyl positions = 380,000 distinct structures. There are also 380,000 potential singly -O-methylated structures; others include acetyl, phosphates, carbamoylates, pyruvates, etc.

A trisaccharide made up of hexoses has 10 possible hydroxyls available for substitution outside of the reducing end. There are 42 unique paired locations (9+8+7+6+5+4+2+1) to substitute 2 sulfates on one trisaccharide. Using 38,000 trisaccharides made of a 3 hexose vocabulary there are **1.6 million possible structures of trihexosides from a library of 3 sugars substituted with 2 sulfates**, or 2 of any single epitope.

There are 84 ways to substitute **two different epitopes** on a trihexose, giving a potential 3.2 million structures for a trisaccharide containing both a sulfate and an acetate, for example.

Using a **20 sugar vocabulary** for trisaccharides there would be 90,000,000 singly sulfated potential structures, 4 X 10^8 potential disulfated trisaccharide isomers, and nearly 10^9 isomers of trisaccharides substituted with one of each of two different epitopes. The isomer numbers are enormously larger for higher oligosaccharides. For a substitution of a single trisaccharide with 3 sulfates, for example, the isomers are at least 240, which you can examine by using 3 rings among 10 fingers.

The carbohydrate side of the chemical information "potential" of the "Carbohydrate Code" has given nature a wide library of structures to choose from (evolutionary potential). In the current state of evolution, perhaps very little of this potential has been used. However, it is not by accident that higher life forms contain more lectins and more numbers of complex carbohydrate structures than simple organisms. In addition to participation in the control of proper protein folding, of targeting proteins glycosylated in the Golgi, there is cell-cell and cell-signal recognition. Carbohydrate structures enhance the complexity of self-recognition systems for the immune system, but one would think that this would give microorganisms a better chance to generate one as a mask.

Examples of biological activities of carbohydrate-protein recognition systems

- The "Selectins", recognition factors for extravasation of leukocytes and therefore mediators of inflammatory responses [14-19]. It is my estimate that more than $300,000,000 has been spent in research on these molecules by the pharmaceutical industry and granting agencies in the past 6 years. Small companies alone have spent a third of this (Glycomed $30,000,000, Cytel $60,000,000). More than 800 papers have been published using the term "selectin" in the title or abstract.

- "NOD" factors are very interesting recognition systems of legumes and nitrogen fixation bacteria in nodulation signals [20-22. These structures are chitin oligomers substituted with sulfates, fucose, acyl groups.

- Recently, it has been noted that pollen tubes follow "glycosylation gradients" in the pistil during the fertilization event in plants[23] (this predicts that a changing array of lectins are expressed on the tip of the growing tubule).

- Cell wall polymers or fragments from fungi stimulate plants to mount host defense systems[24, 25]. Recognition proteins for the polymers must be present in the plants, as yet largely undiscovered.

- N-linked saccharides containing terminal glucose as intermediates are important in the eucaryotic calnexin system that determines whether proteins are properly folded [26, 27].

- Mannose-6-phosphate as a well established targeting system for lysosomal proteins produced in the Golgi is still an active research area [28]. Receptors in this area are also investigated[29.

- Heparin pentasaccharide recognition site for AntithrombinIII [30-38]: More than 1650 papers have been published on this subject since 1985. A large amount of industrial dollars (probably >$100,000,000) have been spent on searching for a more defined anticoagulant subunit structure than the heparin mixture used in medicine. This includes synthesis, a 30-40 step process for which few of the steps are quantitative. In a 30 step synthesis, a 2% loss at each step yields about 50%, 5% loss at each step yields 20%, and the average for a good synthesis may yield >5% losses average at each step. Fragmentation of heparin and purification of active sequences may be a better route, or better yet, controlled chemical modification of N-acetyl heparosan or its oligomers [39].

- Heparin and its interaction with growth factors [40-51] (>1800 papers published on this subject). Intense pharmaceutical research is conducted in this area.

No automated sequencing methods for nanomole amounts of carbohydrates

The use of NMR as a single spectroscopic method to establish a *chemical shift library* for absolute identification of an oligosaccharide is not possible today, even with saccharides as small as 3 sugars. Each trisaccharide from a conservative set of 38,016 isomers using only 3 hexoses would contain 15 ring protons including the reducing end anomeric proton. A proton NMR spectrum library would require a resolution of 38,016 x 15 = 570,240 "different" proton environments within 0.5 ppm where the ring protons are clustered. This would require a resolution of 10^{-6} ppm, (a terahertz instrument) if the line widths were also narrowed concomitantly. It is doubtful that a tenth of this number of lines could be resolved using multi-dimension proton NMR. For ^{13}C, where the lines are thirty times more dispersed, a chemical shift library would need to resolve 38,016 x 18 carbons = 684,288 lines if they happened all to be different. The resolution required for ^{13}C NMR is 2 x 10^{-5} ppm, also in the terahertz range. NMR by itself, therefore, cannot be used to establish a chemical shift library as a stand-alone identification system for trisaccharides and certainly not for larger oligomers.

If a micromole of an unknown saccharide is available, however, NMR is very useful. With only 15 proton lines to resolve in 0.5ppm and with "accidental" overlaps minimized, the use of NOE and 2-D NMR techniques can be used to completely identify the epimers, linkage positions and anomeric configurations. This is very useful for confirmation of synthesis.

In the case of identification of small amounts of biologically active saccharides where often only nanomole quantities are available, microelectrophoresis and glycosyl hydrolase digestion are the most promising current techniques, combined with MALDI Mass Spectrometry. Also, immunochemical methods using lectins and antibodies are becoming more available.

NMR analysis of polysaccharides with repeat units can yield extremely useful information regarding details of structure including *population of derivatization sites, secondary structure and conformation*[39, 52, 53].

No automated organic synthesis methods for carbohydrates

Organic synthesis of oligosaccharides finds itself with barriers, resisting automation for similar reasons, namely that there are too many isomers to establish simple methods. At each step, selective blocking and anomeric mixtures must be dealt with, and the likelyhood of overcoming this problem in a substantial way is of low probability. Unless the product is of extremely high value and tons are necessary, glycosyl transferases may be a more reasonable approach, or combinations of organic and biosynthesis. Glycosyl transferases appear to be very specific when applied in turn to a growing chain. Some glycosyl hydrolases can be used as transglycosylases.

CONCLUSION

Nature has provided a class of compounds which can generate large numbers of recognition factors. More complex organisms require advanced communication among cells, tissues and organs. While all classes of compounds play a role in this communication, carbohydrates are showing an increasing number of interesting biological activities related to molecular recognition. The evolutionary capacity for information in these compounds is very large.

Special thanks to the meeting Organizer Prof. B. Casu, meeting Secretary A. Naggi and the Session Chairman H. Vliegenthart for the opportunity to present this work at the XVIII International Carbohydrate Symposium.

REFERENCES

1 R. Laine. *Glycobiology* **4**, 759-67 (1994).

2 R. Laine. In: *Glycosciences: Status and Perspectives* (eds. Gabius H and Gabius S), pp. 1-15. Chapman & Hall, Weinheim (1997).

3 I. Eggens, B. Fenderson, T. Toyokuni and S. Hakomori. *Biochem Biophys Res Commun* **158**, 913-20 (1989).

4 B. Fenderson, E. Eddy and S. Hakomori. *Bioessays* **12**, 173-9 (1990).

5 K. Koshy and J. Boggs. *J Biol Chem* **271**, 3496-9 (1996).

6 M. Kumar and D. Sarkar. *FEBS Lett* **391**, 17-20 (1996).

7 C. Melito and A. Levy-Benshimol. *Acta Cient Venez* **43**, 312-4 (1992).

8 G. Misevic and M. Burger. *J Biol Chem* **268**, 4922-9 (1993).

9 G. Misevic and O. Popescu. *J Mol Recognit* **8**, 100-5 (1995).

10 L. Sellers and A. Allen. *Symp Soc Exp Biol* **43**, 65-71 (1989).

11 D. Spillmann, J. Thomas-Oates, K. J. van, J. Vliegenthart, G. Misevic, M. Burger and J. Finne. *J Biol Chem* **270**, 5089-97 (1995).

12 R. Stewart and J. Boggs. *Biochemistry* **32**, 10666-74 (1993).

13 Z. Zhu, N. Kojima, M. Stroud, S. Hakomori and B. Fenderson. *Biol Reprod* **52**, 903-12 (1995).

14 L. Lasky. *Annu Rev Biochem* **64**, 113-39 (1995).

15 R. Cummings and D. Smith. *Bioessays* **14**, 849-56 (1992).

16 C. Smith. *Can J Physiol Pharmacol* **71**, 76-87 (1993).

17 A. Varki. *Proc Natl Acad Sci U S A* **91**, 7390-7 (1994).

18 D. Vestweber. *Semin Cell Biol* **3**, 211-20 (1992).

19 J. Whelan. *Trends Biochem Sci* **21**, 65-9 (1996).

20 K. Baureithel, G. Felix and T. Boller. *J Biol Chem* **269**, 17931-8 (1994).

21 S. Ikeshita, Y. Nakahara and T. Ogawa. *Glycoconj J* **11**, 257-61 (1994).

22 R. Poupot, E. Martinez-Romero, N. Gautier and J. Prome. *J Biol Chem* **270**, 6050-5 (1995).

23 H. Wu, H. Wang and A. Cheung. *Cell* **82**, 395-403 (1995).

24 T. Herget, J. Schell and P. Schreier. *Mol Gen Genet* **224**, 469-76 (1990).

25 Y. Okinaka, K. Mimori, K. Takeo, S. Kitamura, Y. Takeuchi, N. Yamaoka and M. Yoshikawa. *Plant Physiol* **109**, 839-45 (1995).

26 W. Chen, J. Helenius, I. Braakman and A. Helenius. *Proc Natl Acad Sci U S A* **92**, 6229-33 (1995).

27 D. Hebert, B. Foellmer and A. Helenius. *Cell* **81**, 425-33 (1995).

28 R. Bresciani and F. K. Von. *Eur J Biochem* **238**, 669-74 (1996).

29 W. Nauseef, S. McCormick and H. Yi. *Blood* **80**, 2622-33 (1992).

30 M. Ragazzi, D. Ferro, B. Perly, G. Torri, B. Casu, P. Sinay, M. Petitou and J. Choay. *Carbohydr Res* **165**, c1-5 (1987).

31 U. Lindahl, L. Thunberg, G. Backstrom, J. Riesenfeld, K. Nordling and I. Bjork. *J Biol Chem* **259**, 12368-76 (1984).

32 U. Lindahl, G. Backstrom, L. Thunberg and I. Leder. *Proc Natl Acad Sci U S A* **77**, 6551-5 (1980).

33 A. Lellouch and P. J. Lansbury. *Biochemistry* **31**, 2279-85 (1992).

34 J. Bae, U. Desai, A. Pervin, E. Caldwell, J. Weiler and R. Linhardt. *Biochem J* **301 (Pt 1)**, 121-9 (1994).

35 D. Atha, A. Stephens, A. Rimon and R. Rosenberg. *Biochemistry* **23**, 5801-12 (1984).

36 B. Casu, P. Oreste, G. Torri, G. Zoppetti, J. Choay, J. Lormeau, M. Petitou and P. Sinay. *Biochem J* **197**, 599-609 (1981).

37 J. Choay, M. Petitou, J. Lormeau, P. Sinay, B. Casu and G. Gatti. *Biochem Biophys Res Commun* **116**, 492-9 (1983).

38 C. van Boekel and M. Petitou. *Angewante Chemistry International Edition English* **32**, 1671-1690 (1993).

39 B. Casu, G. Grazioli, N. Razi, M. Guerrini, A. Naggi, G. Torri, P. Oreste, F. Tursi, G. Zoppetti and U. Lindahl. *Carbohydr Res* **263**, 271-84 (1994).

40 U. Lindahl, K. Lidholt, D. Spillman and L. Kjellén. *Thrombosis Research* **75**, 1-32 (1994).

41 T. Arai, A. Parker, W. J. Busby and D. Clemmons. *J Biol Chem* **269**, 20388-93 (1994).

42 R. Copeland, H. Ji, A. Halfpenny, R. Williams, K. Thompson, W. Herber, K. Thomas, M. Bruner, J. Ryan, D. Marquis-Omer and a. et. *Arch Biochem Biophys* **289**, 53-61 (1991).

43 S. Faham, R. Hileman, J. Fromm, R. Linhardt and D. Rees. *Science* **271**, 1116-20 (1996).

44 M. Klagsbrun, R. Sullivan, S. Smith, R. Rybka and Y. Shing. *Methods Enzymol* **147**, 95-105 (1987).

45 G. Oosta, W. Gardner, D. Beeler and R. Rosenberg. *Proc Natl Acad Sci U S A* **78**, 829-33 (1981).

46 D. Rabenstein, J. Robert and S. Hari. *FEBS Lett* **376**, 216-20 (1995).

47 N. Shao, H. Wang, T. Zhou, Y. Xue and C. Liu. *Life Sci* **54**, 785-9 (1994).

48 A. Strain, G. McGuinness, J. Rubin and S. Aaronson. *Exp Cell Res* **210**, 253-9 (1994).

49 T. Taniguchi, M. Toi and T. Tominaga. *Lancet* **344**, 470 (1994).

50 L. Thompson, M. Pantoliano and B. Springer. *Biochemistry* **33**, 3831-40 (1994).

51 A. Triantos, G. Koliakos, E. Kavoukopoulos, A. Dimitriadou and A. Trakatellis. *Biochem Mol Biol Int* **37**, 737-45 (1995).

52 B. Casu, D. Ferro, M. Ragazzi and G. Torri. In: *Dermatan sulfate proteoglycans, Chemistry, Biology.* (ed. Scott J), pp. 41-43. Portland Press, London (1993).

53 G. Mascellani, L. Liverani, A. Prete, G. Bergonzini, P. Bianchini, G. Torri, A. Bisio, M. Guerrini and B. Casu. *Anal Biochem* **223**, 135-41 (1994).