GLYCOFORUM

Trends in Glycoscience and Glycotechnology
Vol.14 No.77 (May 2002) pp.127–137

A Novel Linear Code® Nomenclature for Complex Carbohydrates

複合糖質のための新しいリニアコード® 命名法

Banin, Ehud; Neuberger, Yael; Altshuler, Yaniv; Halevi, Asaf; Inbar, Ori; Dotan Nir; and Avinoam Dukler*
Glycominds Ltd, 1 Yodfat St., Alon Bldg., Global Park, Lod, 71291, Israel
FAX: 972-8-9181081, E-mail: dukler@glycominds.com

Key Words *: Linear Code, glycomics, nomenclature, carbohydrates, blycobiology.*

Abstract

The Linear Code is a new syntax for representing glycoconjugates and their associated molecules in a simple linear fashion. Similar to the straightforward single letter nomenclature of DNA and proteins, Linear Code presents glycoconjugates in a canonic, compact and practical form while accounting for all relevant stereochemical and structural configurations. It uses a single letter code to represent each monosaccharide and includes a condensed description of the connections between monosaccharides and their modifications, allowing a simple linear representation of these compounds. The new linear syntax enables the implementation of bioinformatics tools for investigation and analysis of glyco-molecules and their biology.

要　約

リニアコードは複合糖質および関連分子を単純な直線的様式で表わす新しい表示法である。ＤＮＡやタンパク質の一文字表示がわかりやすいのと同じ様に、リニアコードはありうる立体化学および立体構造までを含めて、複合糖質を簡潔で実用的な形式で表示する。各単糖に一文字コードをあて、単糖間の結合様式およびいろいろな修飾をも含めて、糖鎖分子を簡単に線状表示できるようになっている。この新しい線状表示法によって、バイオインフォーマティックスのツールを糖分子の分析と生物学的研究へ適用することが可能になる。

A. Introduction

Glycobiology, the study of carbohydrate-containing molecules and their biological activity, was described in the March 2001 special edition of *Scienc*e as a “Cinderella field”. As the Genome project reaches its final stages, the obtained data is confirming that there are only 30,000 genes and there are only small differences between the genomes of different species. The number of native proteins is however much larger, mainly due to post-transcriptional and post-translation modifications of the protein messages. The most common and most diverse post translation modification is protein glycosylation. It has been estimated that more than half of the proteins in nature are glycoproteins (Apweiler *et al.*, 1999). Recent studies have revealed essential roles of carbohydrates in biological processes such as protein folding (Parodi, 2000), protein localization, immunity (Huby *et al.*, 2000), cell proliferation (Zanneta *et al*., 1994), and hormone and growth factor responses (Van den Steen *et al.*, 1998). In addition, many viruses and bacteria use cell-surface carbohydrates to enter cells and subsequently initiate infections (Rossmann *et al.,* 2000; Hooper *et al.*, 2001). The diversity in carbohydrate function makes them exciting new targets for elucidating crucial pathways in a wide range of diseases.

Ａ．はじめに

サイエンスの 2001年 3月の特別号では、糖を含む分子とその生物活性の研究、すなわち糖鎖生物学を「シンデレラ領域」と呼んでいる。ゲノムプロジェクトがほぼ終わり、得られたデータから、遺伝子の総数がわずか 30000程度にすぎず、生物種が違ってもゲノムにはそれほど差がないことが確められた。しかしそのメッセージが転写後および翻訳後に修飾されるために、実際に作られるタンパク質の数ははるかに多い。翻訳後修飾の中でいちばん普遍的でまた多様なものは糖鎖による修飾である。

天然のタンパク質の半分以上は糖タンパク質との見積もりもある(Apweiler *et al*.1999)。さまざまな生命現象において糖質が不可欠な役割をになっていることを最近の研究成果が明らかにしている。タンパク質の立体構造形成(Parodi2000)、タンパク質の局在化、免疫(Huby *et al*.2000)、細胞増殖(Zanneta et al. 2000)、ホルモンや成長因子に対する応答(Van den Steen et al. 1998)などが例としてあげられる。多くのウイルスや細菌が、細胞に侵入して感染を成立させるために細胞表面の糖鎖を利用している(Rossmann et al.2000、Hooper *et al*.2001)。広範な病気の解明に確実につながる路を発見するためにも、糖鎖機能の多様性は研究対象としてきわめて刺激的である。

*Corresponding Author.

Trends in Glycoscience and Glycotechnology
Vol.14 No.77 (May 2002) pp.127–137

Until recently, the field of glycobiology has been largely overlooked. A primary reason has been the extreme complexity and variability of carbohydrates derived from: (a) the types of monosaccharides and modifications present; (b) the types of linkages; and (c) the presence of branching. In addition to creating difficulties in the study of carbohydrates, this structural variability sets up an obstacle for development of a simple and consistent nomenclature. While several recommendations and proposals have been introduced for glycan nomenclature and representation (i.e. IUPAC-IUBMB and Bohne-Lang, *et al.*), the field still suffers from inconsistent use of the designated rules and inconvenient illustration for complex carbohydrates.

The simple linear presentation of amino acids and nucleic acids paved the way for bioinformatics tools, such as databases and homology searches. These tools, which seem trivial today, essentially served as the foundation for genomics and proteomics. In order to develop glycomic tools for databases and bioinformatics, a simple and comprehensive linear representation must first be employed. To meet this need, a new syntax called the Linear Code™ has been developed for representing glycoconjugates and their associated molecules in a simple linear fashion. Similar to the straightforward nomenclature of DNA and proteins, Linear Code presents complex carbohydrates in a compact and practical form while accounting for all relevant stereochemical and structural configurations. This paper describes the novel Linear Code syntax, as well as the symbols and rules used for representation of complex carbohydrates.

B. Carbohydrate representation

There are several established formats for chemical presentation of saccharides. The Fischer and Haworth projections are frequently used. IUPAC-IUBMB has recommended the use of three letter codes for the presentation of monosaccharides and has suggested extended and condensed forms for the presentation of oligo- and polysaccharide chains (Fig 1).

B-1. Linear Code representation of saccharide units

The smallest unit comprising a carbohydrate is the basic saccharide unit (SU). The saccharide unit is composed of four elements: the monosaccharide name, modifications (if any), its anomericity (the α and β configurations of the glycosidic bond) and the position at which it is bound to a given SU. The Linear Code offers a simple way of representing saccharide units and the connections between them.

B-1-1. ***The monosaccharide***

The Linear Code assigns a single letter code to the most common structures of monosaccharides found in vertebrates (Table I). In cases where the monosaccharides are different from the common structure, they are expressed as follows:

- Stereoisomers (D or L) of the common monosaccahrides are indicated with apostrophes: “ ’ “ (MS’).
- Monosaccharides with different ring structure (furanose

糖鎖生物学という領域の重要性はこれまでむしろ見過ごされてきた。その根本的理由として、糖鎖の極端な複雑性と多様性がある。その原因には、 a)単糖の型と修飾、 b)結合様式、 c)分岐、がある。糖質研究をさらに難しくしているのは、構造が多様すぎるために、簡単で合理的な命名法を確立できないことがある。糖質の命名法および表示法についていろいろな勧告や提案がなされてきたが(たとえば IUPAC IUBMB Bohne Lang et al) この領域に依然として残る悩みは、提案された規則に合わない表示が絶えないこと、複合糖鎖を分かりやすく図示できないことの不便さなどである。

アミノ酸や核酸を簡単に線状表示できたことで、データベース検索、ホモロジー検索のようなバイオインフォーマティックスのためのツールは着々と整えられてきた。もはやあたりまえのようなこれらのツールは、ゲノミックスやプロテオミックスの発展に大いに貢献してきた。データベースやバイオインフォーマティックスに役立つグライコミックスのツールを発展させるためには、簡単で完全な線状表示がまず必要である。そこでこの要求に応えて、複合糖質および関連分子を簡単に線状表示できるように、 LinearCodeTM と呼ばれる新しい表示法が作られた。DNAおよびタンパク質についてのわかりやすい命名法と同様に、 LinearCodeは糖質を簡潔で実用的なやり方で、あり得るすべての立体化学的および立体配置まで含めて表現できる。本論文で複合糖質を表示するための記号と規則を含めて、新しい LinearCodeの内容を紹介する。

B 糖の表示法

糖を化学的に表現する方式はいくつもある。 FischerとHaworthの投影法はよく使われる。IUPAC IUBMBは単糖を表示する 3文字コードの使用を勧告し、オリゴ糖鎖や多糖鎖のための総括的な短縮形を提案した(図 1)。

B 1 単位糖の LinearCode表示

糖質を形成する最小の単位は基本的単位糖(SU)である。単位糖は 4つの要素からなる。単糖の名前、修飾(もしあれば)、アノマーの区別(グリコシド結合の と の立体配置)、問題とする SUへの結合位置。 LinearCodeを使えば単位糖およびそれらの間の結合を簡単に表現できる。

B 1 1 単糖

LinearCodeでは脊椎動物にもっとも普遍的な単糖に対して文字のコードをあてる(表 I)。普遍的な構造とは異なる単糖は以下のように表示する。

- 普遍的な単糖の立体異性体(Dまたは L)は引用符号「 」で表す(MS)。
- 普遍的な構造と環構造が違うときは(フラノースとピラ

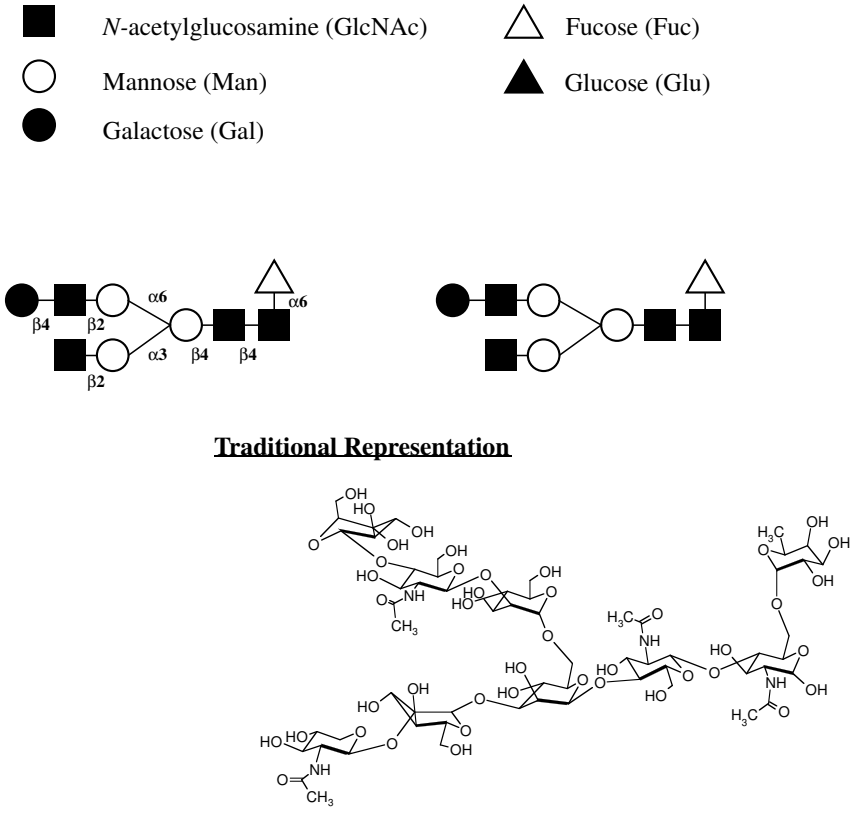


Fig 1. Recommended symbols and conventions for drawing glycan structures compared to the new Linear Code representation. The example used is a typical branched "biantennary" N-glycan with two types of outer termini.

/pyranose) in relation to the common structure are indicated with a caret: " ^ " (MS^).

- Monosaccharides that differ in both stereospecificity and ring structure are indicated with a tilde: " ~ " (MS~).

Example:

D-Galp = A (The most common structure of Galactose)
L-Galp = A'
D-Galf = A^
L-Galf = A~

ノース）挿入記号「^」で表す（MS^）。

・立体異性および環構造の両方が普遍的構造と異なるときは波形「～」で表す。（MS□）。

例：

D□Galp□A（ガラクトースの普遍的な形）
L□Galp□A□
D□Galf□A□
L□Galf□A□

Trends in Glycoscience and Glycotechnology
Vol.14 No.77 (May 2002) pp.127–137

Table I : Linear Codes of common monosaccharide structures (ordered by branch hierarchy).

Trivial Name	Monosaccharide / Core [1]	Linear Code
D-Glc*p*	D-Glucose	G
D-Gal*p*	D-Galactose	A
D-Glc*p*NAc	*N*-Acetylglucosamine	GN
D-Gal*p*NAc	*N*-Acetylgalactosamine	AN
D-Man*p*	D-Mannose	M
D-Neu*p*5Ac	*N*-Acetylneuraminic acid	NN
D-Neu*p*	Neuraminic acid	N
KDN[2]	2-Keto-3-deoxynanonic acid	K
Kdo	3-deoxy-D-manno-2 Octulopyranosylono	W
D-Gal*p*A	D-Galacturonic acid	L
D-Ido*p*	D-Iduronic acid	I
L-Rha*p*	L-Rhamnose	H
L-Fuc*p*	L-Fucose	F
D-Xyl*p*	D-Xylose	X
D-Rib*p*	D-Ribose	B
L-Ara*f*	L-Arabinofuranose	R
D-Glc*p*A	D-Glucuronic acid	U
D-All*p*	D-Allose	O
D-Api*p*	D-Apiose	P
D-Fru*f*	D-Fructofuranose	E

1- all the monosaccharides are in their pyranose form unless otherwise noted.
2- KDN: 3-deoxy-D-glycero-K-galacto-nonulosonic acid.

B-1-2. ***Modifications of the sugar chain***

Modifications are defined as any addition of non-carbohydrate moieties to the basic SU. The modifications are represented by adding square brackets that include the connecting position of the modification to the SU, followed by the modification symbol (Table II) in the form: [#symbol]. For example: D-Glcp with Sulfate (S) in position 3 would be written G[3S]. If there is more than one modification on the same monosaccharide, they are written in numerical order according to their position, within the same brackets. Exceptions include certain monosaccharides with common modifications, for example: *N*-acetylgalactosamine (D-GalpNAc) can be presented by A[2N], but is instead represented with a short two letter code as AN. In the same manner *N*-Acetylneuraminic acid is presented as NN, and *N*-Acetylglucosamine is presented as GN (Table I).

B-1-3. ***Connection to a neighboring MS***

In general, Linear Code uses lower case symbols to represent connecting motifs to the SU such as anomericty, repeat-

B-1-2. 糖鎖の修飾

基本的 SUに付加したすべての非糖質部分を修飾とみなす。修飾は角括弧で表し、そこには修飾の記号(表 II)と結合位置を[＃記号]の形で書く。たとえば D-Glcpの 3位に硫酸基(S)があれば G[3S]。ひとつの単糖に 2つ以上の修飾がある場合は、同じ角括弧の中に、その位置の順番に従って書く。ただしよく知られた特定の修飾単糖は例外とする。たとえば N-アセチルガラクトサミン(D-GalpNAc)は A[2N]と書けるが、2文字コードの ANを使う。同様に N-アセチルノイラミン酸は NNと書き、N-アセチルグルコサミンは GNと書く(表 I)。

B-1-3. 隣接単糖(MS)への結合

原則として Linear Codeは、結合に関する性質、つまりアノマー、繰り返し、環構造などを小文字で表す。隣接する単糖と

Table II : Linear Code of common modifications.

Modification Type	Linear Code
deacetylated *N*-acteyl	Q
ethanolaminephosphate	PE
inositol	IN
methyl	ME
N-acetyl	N
O-acetyl	T
phosphate	P
phosphocholine	PC
pyruvate	PYR
sulfate	S
sulfide	SH
2-aminoethylphosphonic acid	EP

ing and cyclic structures. Two components appear when illustrating the connection between adjacent monosaccharides: the sugar's anomer, and the position at which the sugar is connected to the adjacent sugar. Anomericity is expressed using the letters "a" and "b" – to represent α and β anomers, respectively. These appear immediately following the modification. The connection position will appear after the anomer.

For example:
β-D-Galp(2P)-(1-3)-β-D-Glcp
Would be written as:
A[2P]b3Gb

In cases where a monosaccharide is connected from its first position to a modification and then to another monosaccharide, the modification will be written in square brackets "[ ]" after the anomer, with no number in the brackets (Ab[P]G). If the sugar at the reducing end is in its open form (ol) the letter "o" (in lower case) is added. By convention, carbohydrates are read from right to left. Consistent with this custom, the Linear Code also reads from right to left (i.e. from the reducing end of the carbohydrate).

B-2. Complex carbohydrates

Complex carbohydrates are comprised of sequences of bound saccharide units. Complex carbohydrates range from simple linear forms to highly branched, repeating and cyclic structures. The Linear Code contains a diverse set of rules to account for all possible combinations.

の間の結合を示すときには、2つの要素が必要である。つまり糖のアノマー、隣接単糖に結合する位置である。アノマーの区別は、 と に対して、それぞれ aと bの文字を使う。これらの記号は修飾を表す記号のすぐ後に置く。結合位置はアノマーの後に置く。

たとえば

D Galp 2P 1 3 D Glcp

は以下のようになる。

A [2P] b3Gb

単糖の 1位に修飾があり、その上で他の単糖に結合している場合には、アノマーの記号の後に角括弧 [ ] 内に数字なしで修飾を書く(Ab[P] G)。還元末端の糖が開環構造(ol)であれば、小文字の「 o」をつける。糖質を左から右へと読む習慣にしたがって、LinearCodeでも同様に左から右へ(非還元末端から還元末端へ)と読む。

B 2 複合糖質

複合糖質はつながった単位糖の並びであり、直線的な単純なものから、沢山の枝分かれや繰り返し、環状構造などさまざまである。LinearCodeはありうる組み合わせすべてに対応できるように、多様な規則を含んでいる。

B-2-1. ***Linear complex carbohydrates***

A linear complex carbohydrate refers to an unbranched, non-cyclic string of saccharide units.

Example:

α-L-Fucp-(1-4)-β-D-GlcpNAc-(1-3)-β-D-Galp-(1-4)-β-D-Glcp

Would be written as:

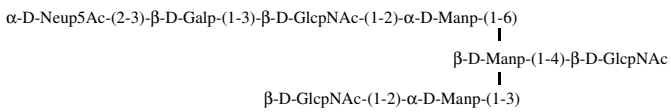
Fa4GNb3Ab4Gb

B-2-2. ***Branch points***

The major complexity of glycan presentation is due to the variability in complex carbohydrates' branching structure, ranging from linear to branched to polymeric. The Linear Code allows linear representation of branched carbohydrates by expressing the branches in parentheses "( )". Deciding which chain is the branch and which is the backbone is accomplished by using two basic rules:

1. When the monosaccharides commencing each chain are identical, the chain connected to the higher position is considered the branch. This rule is also applicable in cases where different modifications exist (when the modified form is not in the monosaccharide hierarchy table).

For example:



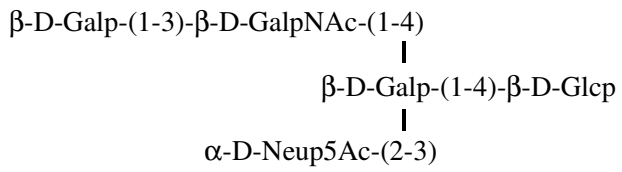
Would be written as:

GNb2Ma3(NNa3Ab3GNb2Ma6)Mb4GNb

Since both branches are commencing with Mannose, the branch chain vs. backbone chain is designated as follows: the chain beginning with the Mannose connected to the higher position (6 position in example) is the branch chain.

2. When the saccharide units at the branch point are different, the single letter code table (Table I) is utilized. The monosaccharides in Table I are organized according to hierarchy that was empirically determined according to the frequency in which certain sugars appear at the branch node, in order to normalize the data. The chain beginning with the lower MS in the hierarchy table (thus the more rare SU), is designated the branch chain and will be written inside the parenthesis "( )". Concurrently, the chain beginning with the higher MS rank is designated the backbone chain. Modifications do not change the hierarchy of the MS except for the modified MSs existing in the table itself.

For example:



B-2-1. 線状の複合糖質

単位糖が分岐なしに、また環も作らずにつながっている複合糖質。

例

α-L-Fucp-(1-4)-β-D-GlcpNAc-(1-3)-β-D-Galp-(1-4)-β-D-Glcp

は以下のように表示される。

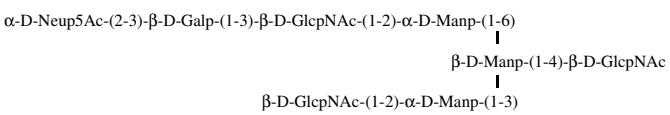
Fa4GNb3Ab4Gb

B-2-2. 分岐の位置

グリカンの表示が複雑になる最大の理由は、直鎖構造から分岐構造、さらには高分子化といった複合糖質特有の多様な分岐構にある。Linear Codeでは枝を括弧( )内に書くことによって、分岐糖鎖であっても線状に表示できる。次の 2つの規則で、どの鎖が枝で、どの鎖が幹かを判定する。

1. 各糖鎖の最初の単糖が同じ場合には、結合位置の数字の大きい方を枝と考える。この規則は修飾のされ型が異なる場合 (その修飾が単糖の階層表にない場合)にも適用される。

例



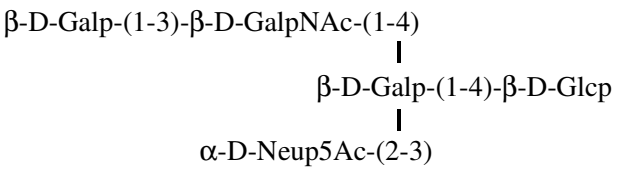
は下のように表示される。

GNb2Ma3(NNa3Ab3GNb2Ma6)Mb4GNb

いずれの糖鎖もマンノースから始まっているので、枝と幹の区別は以下のようになる。数字の大きい位置(例では 6位)に結合したマンノースから始まる糖鎖を枝とする。

2. 分岐点での単位糖が異なる場合、 1文字コード表(表 I)を利用する。表 Iではデータを規格化するために、それぞれの単糖が分岐点に出現する頻度をもとにして、経験的に決めた単糖の階層を示している。低位のMS(出現頻度の低い単位糖)から始まる糖鎖を枝とみなし、括弧( )の中に書く。同時に高位のMSから始まる糖鎖を幹とみなす。表に含まれている修飾単糖以外には、修飾で階層が変更されることはない。

例

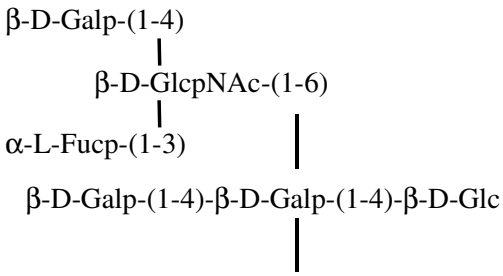


Would be written as:

Ab3ANb4(NNa3)Ab4Gb

Since NN is lower in hierarchy than AN (see Table I), the chain beginning with NN is considered the branch chain.

When the complex carbohydrate contains more than two branches, the decision of branch / backbone is initially determined by the aforementioned hierarchy rules, and then according to the monosaccharide unit position. The following example shows the rules for determining the Linear Code of a branched carbohydrate structure.



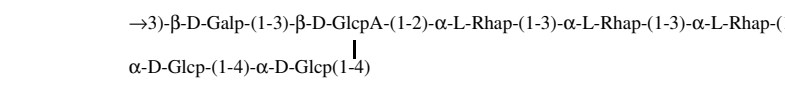
Would be written as:

Ab4(GNb4GNb3)(Ab4(Fa3)GNb6)Ab4Gb

Starting from the reducing end, there are two branches and a backbone. The first branch chain, (i.e. the chain commencing with GNb6) was determined according to the hierarchy table (Gal vs. GlcpNAcp – see Table I) and the higher connection position (GNb6 vs. GNb3). Since the first branch chain is branched again (Fa6 versus Ab4) a nested branch point is added based on the hierarchy table (i.e. Fa3 is branched). The second branched chain (GNb4GNb3) is then added and finally the backbone - Ab4.

B-2-3. ***Repeating and cyclic units***

In some cases complex carbohydrates will contain cyclic or repeating units. In the Linear Code, a cyclic motif is expressed using the letter “c”. Repeating units are expressed inside parentheses “{n }”, where ‘n’ represents the number of repeats. For example, cellulose, which is a polymer of D-Glucose residues joined by β-1,4 linkages, would be written as: {nGb4}. If the repeating units are not connected ‘head to tale’, the monosaccharide at which the unit is connected is marked between two dashes “- -“. For example the capsular polysaccharide of *Klebsiellae* serotype K79 (Guy *et al*., 1985) has the following repeating unit:



Would be written as:

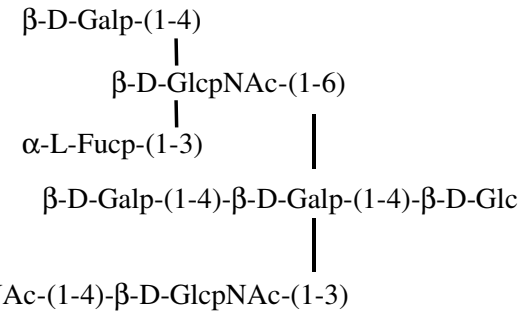
{nGa6Ga4(-Ab3-)Ub2Ha3Ha3Ha3}

は次のように表示される。

Ab3ANb4(NNa3)Ab4Gb

NNが ANよりも低い階層(表 I)なので、NNではじまる鎖が枝とみなされる。

複合糖鎖が 3本以上の枝をもつとき、枝と幹の判定は先ずこれまでと同じ規則、次いで単糖単位の位置で考える。多数の枝をもつ糖鎖を LinearCodeでどう書くかの例を下に示す。



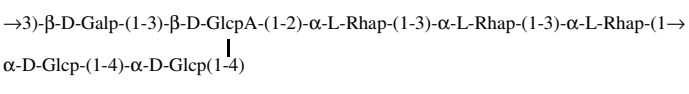
は次のように表示される。

Ab4(GNb4GNb3)(Ab4(Fa3)GNb6)Ab4Gb

還元末端から出発して 2本の枝と 1本の幹がある。第 1の枝(GNb6から始まる鎖)は階層表(Galと GlcNAcp、表 I参照)と、結合位置の数字の大きさ(GNb6と GNb3)にしたがって決まる。第 1の枝はさらに分岐するので(Fa6と Ab4)、階層表にもとづいて分岐点につけ加えられる(Fa3が分岐)。最後に第 2の分岐鎖(GNb4GNb3)が幹の Ab4につけ加えられる。

B-2-3. 繰り返し単位と環状単位

複合糖質には環状あるいは繰り返し単位が含まれる場合がある。LinearCodeでは環状モチフを「c」で表す。繰り返し単位は{n}のように、括弧の内部に繰り返し回数 nを書いて表す。たとえばセルロースは D-グルコース残基が β-1,4結合したポリマーなので、{nGb4}のように表す。繰り返し単位が「頭から尾」のようにつながっていない場合は、その単位がつながっている単糖を 2つのダッシュではさむ。たとえば Klebsiellaeの血清型 K79の莢膜多糖(Guy et al., 1985)は次のような繰り返し単位を持っている。



は以下のように表示される。

{nGa6Ga4(-Ab3-)Ub2Ha3Ha3Ha3}

B-2-4. ***Glycoconjugates***

A saccharide unit is often connected through its reducing end to various non-carbohydrate moieties. The Linear Code divides this type of connection into three groups: amino acid sequences, lipid moieties and 'other' molecules. The representations of these groups are as follows:

• Amino acid sequences are written after a semicolon "**;**" using the amino acid single letter code. For example: α-D-Glc bound to Asn-Tyr-Ser-Cys would be written as: Ga;NYSC. In cases were the SU is bound to an amino acid in the middle of the sequence (Ser in the aforementioned example) the amino acid is marked using "- -" . So the Linear Code would be Ga;NY-S-C.

• Lipid moieties are written after a colon "**:**", using the Linear Code representation for lipids (Table III). For example:β-D-Glc bound to Ceramide would be written as: Gb:C.

• Other glycosides are written after the number symbol "**#**" using its complete name. For example: β-D-GlcpNAc-(1-3)-β-D-Galp-(1-1)-4-Trifluoroacetamidophenol would be written as: GNb3Ab#4-Trifluoroacetamidophenol

B-3. Unknown and Uncertain elements

Due to the structural diversity of carbohydrates, and the fact that determination of oligosaccharide sequences are far from routine, there may be one or more components in a saccharide unit or in a complex carbohydrate that are unknown or uncertain. The Linear Code accounts for unknown or uncertain components in the following manner.

B-3-1. ***Unknown components of the saccharide unit***

If only one component of a SU is unknown, a single question mark "?" should be used. For example AN?3G, represents a SU with an unknown anomer type (α or β), if the connection position is also unknown, the Linear Code would be written as AN??G. When an entire saccharide unit in the complex carbohydrate is unknown, the asterisk character " * " is used. For example: the Linear Code ANb3*A contains 3 saccharide units, but the identity of the middle SU is unknown.

B-2-4 複合糖質

糖質は還元末端で糖質以外の分子にも結合する。Linear Codeではこうした結合を、アミノ酸配列、脂質部分、その他 3つのグループに分ける。それぞれを以下のように表示する。

・アミノ酸配列はセミコロン「；」の後に 1文字コードを使って書く。たとえば、Asn-Tyr-Ser-Cysに □-D-Glcが結合している場合、Ga;NYSCと書く。SUがアミノ酸配列の中間に結合している場合(上の例でたとえばSerに結合)、そのアミノ酸を「 - 」でマークする。したがって Linear Codeは Ga;NY-S-Cとなる。

・脂質部分はコロン「：」の後に、脂質の Linear Code(表 III) に従って書く。たとえば □-D-Glcがセラミドに結合していれば Gb:Cのように書く。

・その他のグリコシドは記号「＃」の後に、完全な名前を書く。たとえば

□-D-GlcpNAc-(1-3)-□-D-Galp-(1-1)-4-Trifluoroacetamidophenol は GNb3Ab#4-Trifluoroacetamidophenolと書く。

B-3 未知あるいは不確定な要素

糖質の構造が多種多様性であり、オリゴ糖鎖の配列を機械的に決定できるようになるのはまだまだ先のことなので、ある糖鎖あるいは複合糖鎖について、いくつかの構成成分が未知あるいは不確かということもありうる。このような場合 Linear Codeでは以下のようにして表示する。

B-3-1 糖質の中の未知の成分

もしひとつの構成成分だけが未知の場合、1個の疑問符「？」をつける。たとえば AN?3Gはアノマーに関して□か□かがわからないことを示す。もしも結合位置も未知の場合は Linear Codeでは AN??Gと書く。複合糖質内の 1つの単位糖が完全に未知の場合はアステリスク「＊」を使う。たとえば Linear Code ANb3*Aは 3つの単糖からなることはわかっているが、中央の SUは同定されていない。

Table III : Lipid moieties in Linear Code.

Trivial name	Full name	Linear Code
Cer	Ceramide	C
Sph	Sphingosine	D
IPC	Inositolphosphoceramide	IPC
DAG	Diacylglycerol	DAG

B-3-2. ***Uncertain components of the saccharide unit***

When two possibilities are given for the identity of a saccharide unit element, a slash "/" is used to separate the two options. For example: ANb3/4 states that the binding position can either be carbon number 3 or 4 of the neighboring saccharide unit.

B-3-3. ***Uncertain saccharide units***

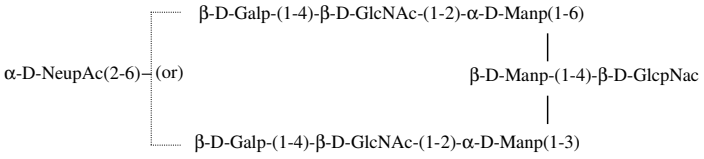
When two options are given for the identity of a complete saccharide unit, two slashes "//" are used to separate the two options. For example: the Linear Code Ab4//Ga2Aa3 states that the actual structure can either be Ab4Aa3 or Ga2Aa3.

It is important to emphasize that both the slash and question mark characters may be used several times in one complex carbohydrate.

B-3-4. ***Uncertain connection site(s) of a saccharide unit to the complex carbohydrate***

When analyzing branched complex carbohydrates, there are often uncertainties in the connection site of a saccharide unit (or a sequence of saccharide units) to a given complex carbohydrate. In other words, we may not know whether a saccharide is attached at point A or point B. Therefore we have designated a Linear Code rule to describe this uncertainty.

An uncertainty is labeled by a variable. The variable is written as two characters, a percentage symbol and the variable index number. For example: 1% and 2% represent two separate uncertainties. The full description (i.e. complete saccharide unit names) of each of the possibilities, "index%", are placed at the end of the Linear Code, after the vertical bar symbol "|". The "|" symbol separates between a given Linear Code and its possibilities. For example:



Would be written:

NNa6=1%|1%Ab4GNb2Ma3(1%Ab4GNb2Ma6)Mb4Gb

The example shows a structure where the uncertainty is in the exact connection site of NeupAc(2-6) at the non-reducing ends of the complex carbohydrate (i.e. to the Manα1-3 or Manα1-6 branch). Since there is only one "uncertain" saccharide unit (NNa6), the Index number is 1; and thus,"1%" is added to both non-reducing ends (Manα1-6 and Manα1-3). In addition the "NNa6=1%|" at the end of the Linear Code is added as a footnote in order to describe the uncertain saccharide unit.

B-3-2. 単位単糖が不確実な場合

単位単糖が何であるかについて 2つの可能性がある場合は、スラッシュ「/」で 2つの可能性を分けて示す。たとえば ANb3/4は隣の単糖と結合する炭素が 3位または 4位であることを示す。

B-3-3.

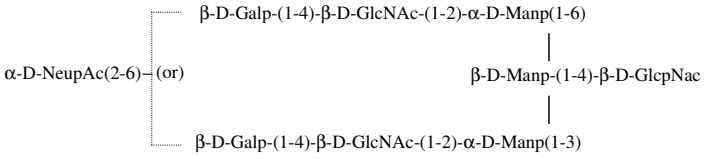
糖鎖全体に対して 2つの可能性がある場合は、2本のスラッシュで 2つの可能性を分けて示す。たとえば LinearCode Ab4//Ga2Aa3は、実際の構造が Ab4Aa3か Ga2Aa3のどちらかであることを示す。1つの複合糖質に対して、疑問符、スラッシュともに複数回も使えることを強調しておく。

B-3-4. 複合糖質に対する糖鎖の連結位置が不確実な場合

分岐している複合糖質を解析する場合、単糖あるいは糖鎖の連結位置が不確実なことがしばしばある。言い換えれば、ある単糖が A点と B点のどちらに連結しているかを明確にできない場合がありうる。このような場合に対する LinearCodeも考案してある。

ひとつの不確実性をひとつの変数で標識する。変数は 2つの文字、すなわちパーセント記号と数字、の組み合わせからなる。たとえば 1%と 2%は、異なる 2つの不確実性があることを示す。「数値%」を、ありうる構造を完全に記載(完全な糖単位名)した LinearCodeの端に、縦線「|」の後に書く。縦線「|」は注目する LinearCodeとそれがつながる可能性との間を分ける。

たとえば



は次のようになる。

NNa6=1%|1%Ab4GNb2Ma3(1%Ab4GNb2Ma6)Mb4Gb

この例は NeupAc(2-6)が複合糖質のどちらの枝(Man α1-6枝と Man α1-3枝)の非還元末端に連結しているかが不確実だということを示している。不確実な単位単糖(NNa6)は 1種類しかないので、指標数字 = 1で、非還元末端の両方(Man α1-6枝と Man α1-3枝)に 1%をつける。また不確実な単位単糖を示すために、注釈として NNa6=1%|を LinearCodeの端につける。

C. Conclusions

Glycobiology has steadily grown in recent years as a promising field for discovery of novel medicines. As the scientific community seeks new insight into the fascinating world of glycans, new bioinformatics tools must be developed in order to facilitate and analyze the growing amount of data. A consensus linear presentation of carbohydrates is essential for the development of these tools. The Linear Code enables, for the first time, the development of carbohydrate bioinformatics tools. Moreover complex carbohydrates can now be stored in databases in a comprehensive and user-friendly manner. Substantial progress can be made in the availability of primary and added-value databases, the development of algorithms, and network information services for carbohydrate analysis. These services can be used for glycan homology searches, elucidation of complex carbohydrate biology, and inter-species comparison of glycans to identify unique and common structures.

We have already developed several glycomic tools that are based on the new Linear Code syntax. The Glyder algorithm allows for the first time structural comparison of glycans. The Glycomics Database (www.glycomics.com) is an advanced database that compiles information about glyco-conjugates. The Linear Code can make a key contribution to the organization and analysis of the massive amount of glycan information available, allowing in-depth investigation of glyco-molecules and their biology.

Acknowledgments

We thank Prof. Nathan Sharon for his assistance during the preparation of the manuscript. The Linear Code is a Registered Trademark of Glycominds Ltd.

C. 結語

糖鎖生物学は新しい医薬を発見しうる可能性を秘めた領域として、近年、着実に成長してきた。科学界が魅力的な糖質の世界における新しい視点を探っており、増大するデータを効率よく解析するために、新しいバイオインフォーマティックスのツールを発展させる必要がある。こうしたツールを発展させるためには、合意された糖質の線状表示法が不可欠である。Linear Codeは糖質のバイオインフォマティックスツールを発展させうる最初のものである。また網羅的で利用しやすく、複合糖質をデータベースに蓄積できるようになっている。基本的および付加価値のあるデータベースの利用、アルゴリズムの進歩、糖質解析のための情報サービスネットワークによって、今後の着実な発展が期待できる。こうしたサービスは糖質のホモロジー検索、複合糖質の生物機能の解明、生物種間での共通構造や特異構造の比較などにも利用できる。

われわれはすでに新しいLinear Code書式にもとづいたいくつかのグライコミック研究のためのツールを開発している。GlyderTMアルゴリズムは初めてグリカンの構造比較を可能にした。Glycomics Database（www.glycomics.com）は糖質との複合体となった分子の情報を編集した発展したデータベースである。Linear Codeは糖質に関して得られた多大な情報の組織化と解析に大きく貢献し、糖質分子とその生物学に関する研究を深めるであろう。

帝京大学薬学部
笠井　献一訳

References

Apweiler, R., Hermjakob, H. and Sharon N. (1999) Biochim. Biophys. Acta **1473**, 4–8
Bohne-Lang, A., Lang, E., Forster, T., and von der Lieth, C.W. (2001) Carbohydr. Res. **336**, 1–11
Guy, G., Dutton, S. and Lim, V. (1985) Carbohy. Res. **144**, 263–267
Hooper, L. and Gordon, J. (2001) Glycobiology **2001,** 1R–10R
Huby, R., Dearman, R. and Kimber, I. (2000) Toxicol. Sci. **55**, 235–246
Parodi, A. (2000) Annu. Rev. Biochem. **69**, 69–93
Rossmann, M., Bella, J., Kolatkar, P., He, Y., Wimmer, E., Kuhn, R., and Baker, T. (2000) Virology **269**, 239–247
Van den steen, P., Rudd, P., Dwek, R. and Opdenakker, G. (1998) Curr. Rev. Biochem. Mol. Biol. **33**, 151–208.
Zanetta, J., Badache, A., Maschke, S., Marschal, P. and Kuchler, S. (1994) Histol. Histopathol. **9**, 385–412.

Profile of the Authors

Ehud Banin joined Glycominds in early 2000, and has been the Director of Data Research since May 2001. Aside from his management responsibilities, Mr. Banin oversees the infrastructure of the Glycomics Database, and the implementation of various glyco-bioinformatics tools associated with this database. Mr. Banin received his B.Sc. from the Technion Israel Institute of Technology in 1996, and is expected to receive his Ph.D. in Biotechnology from the Tel Aviv University during June, 2002. His doctorate study focused on the characterization of pathogenic bacterial organisms. Mr. Banin has 12 scientific publications in various aspects of microbiology.

★★★★★★★★★

Dr. Dukler received his Ph.D. in Biotechnology from Tel Aviv University in 1998, where his research focused on various bioprocessing techniques including affinity-based in situ product removal. In 1999 Dr. Dukler co-founded Glycominds and has since held the position of Glycominds Chief Executive Officer. Dr. Dukler initiated and led the development of Linear Code, Glycomics Database and GlyoChip® technologies, the company's core technology platforms. His major research interests include bioinformatics and data mining. He has filed 12 U.S. patents, and has written two pioneering publications in the field of affinity based in situ product removal.