Counting glycans revisited

Sebastian Böcker and Stephan Wagner

Lehrstuhl für Bioinformatik, Friedrich-Schiller-Universität Jena, Ernst-Abbe-Platz 2, Jena, Germany,
sebastian.boecker@uni-jena.de
and
Department of Mathematical Sciences, Stellenbosch University, Private Bag X1, Matieland 7602, South Africa,
swagner@sun.ac.za

Abstract We present an algorithm for counting glycan topologies of order n that improves on previously described algorithms by a factor n in both time and space. More generally, we provide such an algorithm for counting rooted or unrooted d-ary trees with labels or masses assigned to the vertices, and we give a "recipe" to estimate the asymptotic growth of the resulting sequences. We provide constants for the asymptotic growth of d-ary trees and labeled quaternary trees (glycan topologies). Finally, we show how a classical result from enumeration theory can be used to count glycan structures where edges are labeled by bond types. Our method also improves time bounds for counting alkanes.

Keywords: Counting glycans, counting chemical structures, counting trees, Pólya's enumeration theorem, algorithms

Mathematics Subject Classification: 92E10

1 Introduction

Glycans are — besides nucleic acids and proteins — the third major class of biopolymers. Glycans may occur attached to proteins or lipids, or as free oligosaccharides in the cell plasma. They are believed to play an important role in cell growth and development, tumor growth and metastasis, immune recognition and response (Raman et al., 2005), and even the allergic reaction to white wine (Palmisano et al., 2010). The elucidation of glycan structure remains one of the most challenging tasks in biochemistry. Like metabolites, but unlike proteins, the structure of glycans cannot be directly inferred from the genome sequence of an organism.

The building blocks of glycan polymers are simple sugars (monosaccharides). Since monosaccharides can have up to five linkage sites, glycans can be assembled in a tree-like structure, making their primary structure considerably more complex than that of proteins. Two monosaccharides can be concatenated by a glycosidic bond that is formed between the anomeric carbon of one monosaccharide and a hydroxy group of another: Chemically speaking, the hemiacetal group of one monosaccharide reacts with the alcohol group of the other monosaccharide, releasing water. Each monosaccharide has one "in-link" (the anomeric carbon) and usually four "out-links" (carbon hydroxy groups). The *reducing* end of a glycan bears a free anomeric center that is not engaged in a glycosidic bond (Varki et al., 2009). It is still being referred to as reducing end in case the monosaccharide is in fact linked to a serine or threonine of a protein, or a lipid.

We can model the *glycan topology* as a rooted tree $T = (V, E)$, where the root of the tree is the monosaccharide that forms with the reducing end of the glycan. Tree vertices are labeled with monosaccharides from a fixed alphabet Σ, where Σ depends on the biological background of the experiment. Every vertex has an out-degree of at most four, because each monosaccharide has at most five linkages.[1] A glycan topology does not contain information regarding which of the "out-links" (carbon hydroxy groups) is actually used for an edge. A *glycan structure* reports, for every edge, the type of bond. This is done by numbering the carbons of the monosaccharide, so that links can be labeled, say, "1–3".

Mass spectrometry is one of the predominant experimental techniques for glycan analysis, and allows us to determine the mass of a glycan molecule with high accuracy. Many approaches for deriving glycan topologies from mass spectrometry data are based on enumerating all topologies of a particular mass, and comparing a simulated spectrum for each topology against the measured data (Gaucher et al., 2000; Ethier et al., 2003; Goldberg et al., 2006). Counting glycan topologies and, in particular, the number of topologies of a certain mass M allows us to estimate the algorithm's running time before actually running it, or to check its general applicability. A typical range for this type of application is $M = 0, \ldots, 3000$. In accordance with computational mass spectrometry literature, we will speak about the "mass" (and not the "weight") of combinatorial structures. If we want to take into account decimal places of monosaccharide masses in our calculations, then the mass range of interest is much larger: With two decimal places, the mass range increases to $M = 0, \ldots, 300000$.

In this paper, we consider the question of counting glycan topologies and glycan structures. We stress that we concentrate solely on counting the combinatorial representations of these objects; we do not take into account restrictions such as, "can the glycan topology be realized in 3D space?" Counting chemical structures has a long and interesting history. The most famous example is probably the problem of counting alkanes (saturated acyclic carbohydrates) (Henze and Blair, 1931), whose history starts with Cayley (1881). Amongst other things, it inspired Pólya to develop his enumeration theory (Pólya, 1937) which applies to many other combinatorial questions as well. The first computer implementation is due to Trinajstić et al. (1983). A more recent contribution to the subject is Rains and Sloane (1999). The book by Harary and Palmer (1973) is a standard textbook reference.

A naive application of Pólya's enumeration theorem (Pólya, 1937) allows us to count the number of glycan topologies with n vertices (monosaccharides) in $O(n^4)$ time and $O(n)$ space, see eq. (2) below. Böcker et al. (2011) presented an algorithm for counting glycan topologies in $O(n^3)$ time and $O(n^2)$ space, and for counting all d-ary, vertex-colored trees in $O(d^2n^3)$ time and $O(dn^2)$ space. They also presented an algorithm for counting all glycan topologies of a given mass M in $O(|\Sigma|\, M + M^3)$ time and $O(M^2)$ space. It should be mentioned here that we count the number of operations as "running time" and the number of integer values that have to be stored as "space". Since, however, the numbers we want to compute grow exponentially with

[1] In this paper, we do not consider the comparatively rare case of monosaccharide which do not have exactly four carbon hydroxy groups.

n as well, the actual space requirement is a factor n higher, and even the output itself necessarily needs $\Theta(n)$ bits. This phenomenon also applies to the arithmetic operations involved, as we have to multiply large integers.

In modern terminology, determining the number of things is usually called "counting", whereas "enumerating" refers to constructing all objects with a particular property. But in "classical" terminology, "to enumerate" was often used as a synonym for "to count". In this paper, we concentrate solely on counting certain structures, and in accordance with classical terminology we will use "to enumerate" solely as a synonym of "to count". We note in passing that our methods also allow us to enumerate all structures that have a particular property, provided that it can be modeled in terms of generating functions (for example, structures where the number of monosaccharides of a certain type is prescribed).

Our contributions. In this paper, we address several questions in the context of counting glycan topologies. We present improved algorithms for counting glycan topologies as well as general (vertex-colored) trees with bounded out-degree. We show how classical techniques from combinatorial analysis allow us to count all glycan topologies in $O(n^2)$ time and $O(n)$ space, thus improving upon previous results by a factor n both for running time and space. We modify our approach so that glycan topologies of a particular mass can be counted. We generalize our approach to count all d-ary trees with vertex masses from a given multiset Σ in $O(dM^2 \log d + |\Sigma|\, M)$ time and $O(dM)$ space where M now denotes the total mass. This again improves upon previous results by a factor M in both cases. For that, we use a combinatorial trick to avoid the superpolynomial growth associated with the partition number of d, that is, the number of ways of writing d as a sum of positive integers. We give approximation formulas for the number of glycan topologies for an exemplary monosaccharide alphabet, both for fixed number of vertices n and fixed mass M. We also show how to calculate the constants for the asymptotic growth of these recurrences with high accuracy; while this is well-known to experts in analytic combinatorics, there seems to be no "recipe" available how to do this in practice (Harary et al. (1975) provide a more general algorithm for tree enumeration, but it is not exactly a simple recipe either). Finally, we show how to count glycan structures where the type of each glycosidic bond is known, using a well-known characterization from combinatorics. Our method also allows us to count the number of alkanes (Henze and Blair, 1931) in $O(n^2)$ time and $O(n)$ space which, to the best of our knowledge, is the fastest algorithm for this question, too.

2 Counting glycan topologies with n vertices

Let us start with the most basic case: enumeration of glycan structures by number of vertices (monosaccharides), without labeling vertices. This is equivalent to the enumeration of non-isomorphic rooted trees whose interior vertices all have out-degree four. The classical approach to this problem involves generating functions and Pólya's enumeration method (Pólya, 1937): Let $r[n]$ be the number of non-isomorphic rooted trees with n internal vertices all of which have exactly four children. If $R(x)$ is the associated generating function, i.e.,

$$R(x) := \sum_{n \geq 0} r[n] x^n,$$

then

$$R(x) = 1 + xZ(S_4, R(x)) = 1 + \frac{x}{24}\left(R(x)^4 + 6R(x)^2R(x^2) + 3R(x^2)^2 + 8R(x)R(x^3) + 6R(x^4)\right). \quad (1)$$

Here, $Z(S_k)$ is a so-called *cycle index* (see (Harary and Palmer, 1973)), defined as

$$Z(S_k) := Z(S_k, s_1, s_2, \ldots, s_k) = \frac{1}{k!} \sum_{\pi \in S_k} \prod_{j=1}^{k} s_j^{c_j(\pi)},$$

where $c_j(\pi)$ is the number of cycles of length j in a permutation π. Cycle index $Z(S_k, R(x))$ is obtained from $Z(S_k)$ by replacing s_j with $R(x^j)$ for every j.

The terms in (1) can be interpreted, in a somewhat simplistic way, as follows: the first term stands for the case of the tree without internal vertices (i.e., a single leaf). All other trees consist of a root and four

branches, each of which is again a rooted tree. These quadruples of rooted trees are enumerated by the generating function $R(x)^4/24$, where the denominator makes sure that the $4! = 24$ different permutations are only counted as one tree. Trees with two identical (and two arbitrary) branches are counted only with a factor $12/24 = 1/2$ now, since there are only twelve different permutations. Hence, the correction term $R(x)^2R(x^2)/4$ is needed, where each such tree is counted twice due to the $1!\cdot 1!\cdot 2! = 2$ possible permutations. Likewise, trees with two times two identical branches are counted with a factor of $6/24 = 1/4$ plus $1/4$, and the term $R(x^2)^2/8$ corrects this as there are $2!\cdot 2!$ permutations here. Finally, the term $R(x)R(x^3)/3$ takes account of trees with three identical branches, and $R(x^4)$ of those with four identical branches.

While the above approach is very classical, it seems that there is almost no literature on how to turn such functional equations into efficient algorithms for computing the coefficients $r[n]$. Naively, one could translate the functional equation (1) into a recursion for the coefficients $r[n]$:

$$\begin{aligned} r[n] = &\frac{1}{24}\sum_{j=0}^{n-1}\sum_{k=0}^{n-j-1}\sum_{\ell=0}^{n-j-k-1} r[j]r[k]r[\ell]r[n-j-k-\ell-1] \\ &+ \frac{1}{4}\sum_{k=0}^{\lfloor (n-1)/2\rfloor}\sum_{\ell=0}^{n-1-2k} r[k]r[\ell]r[n-1-2k-\ell] \\ &+ \begin{cases} \frac{1}{8}\sum_{k=0}^{(n-1)/2} r[k]r[(n-1-2k)/2] & \text{if } n-1 \text{ is even,} \\ 0 & \text{otherwise} \end{cases} \\ &+ \frac{1}{3}\sum_{k=0}^{\lfloor (n-1)/3\rfloor} r[k]r[n-1-3k] + \begin{cases} \frac{r[(n-1)/4]}{4} & \text{if } 4 \mid n-1, \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{2}$$

This yields an algorithm that requires $O(n^4)$ time and $O(n)$ space to compute $r[n]$. It is not hard to improve upon this approach by storing additional values, namely the coefficients of $R(x)^2$, $R(x)^3$, and $R(x)^4$: Then all recursions only involve single summations, resulting in a total time of only $O(n^2)$. Let $r_i[n]$ denote the coefficients of $R(x)^i$ for $i = 2, 3, 4$ ($r_1[n]$ would simply be $r[n]$). Then we have

$$r_2[n] = \sum_{k=0}^{n} r[k]r[n-k], \quad r_3[n] = \sum_{k=0}^{n} r[k]r_2[n-k], \quad r_4[n] = \sum_{k=0}^{n} r[k]r_3[n-k], \tag{3}$$

and

$$\begin{aligned} r[n] = &\frac{1}{24}r_4[n-1] + \frac{1}{4}\sum_{k=0}^{\lfloor (n-1)/2\rfloor} r[k]r_2[n-1-2k] + \frac{1}{8}\sum_{k=0}^{(n-1)/2} r[k]r[(n-1-2k)/2] \\ &+ \frac{1}{3}\sum_{k=0}^{\lfloor (n-1)/3\rfloor} r[k]r[n-1-3k] + \frac{r[(n-1)/4]}{4} \end{aligned} \tag{4}$$

from (1), where we interpret $r[m] = 0$ if m is not an integer. This saves a factor of n^2 in the algorithmic complexity. This is similar to the convolution trick in (Böcker et al., 2011), but the running time reported here is a factor n better. We reach:

Proposition 1. *The exact number of rooted trees with maximum out-degree four and n vertices can be computed in $O(n^2)$ time and $O(n)$ space using recurrences* (3) *and* (4).

3 Counting glycan topologies with mass M

In this section, we generalize our idea slightly to the case that each vertex bears a certain mass and we are counting trees with respect to their total mass. Throughout this section, we assume that we are given a multiset $\Sigma = \{m_1, m_2, \ldots, m_J\}$ of possible (integer) masses, which do not have to be distinct. We assign a mass from Σ to each internal vertex, where artificial leaves (which we added to obtain trees whose internal vertices all have out-degree four) get mass zero. The total mass of a tree is the sum of masses of its vertices,

and we are interested in the number of trees with a given mass M. This covers also the case from the previous section, where we simply assume $\Sigma = \{1\}$ and $M := n$. Furthermore, this allows us to count glycan trees with n vertices where each vertex is labeled with a certain color (monosaccharide): In this case, we assume that $\Sigma = \{1, \ldots, 1\}$ is a multiset of appropriate cardinality and, again, set $M := n$.

The approach of the previous section can easily be generalized to this new situation: To allow for a uniform presentation of our calculations, assume that $n = M$ is the mass we want to decompose. Let $r[n]$ now denote the number of rooted trees with total mass n whose internal vertices all have out-degree four, and let

$$R(x) := \sum_{n \geq 0} r[n] x^n$$

be the associated generating function. Then we have, in analogy to (1), the functional equation

$$\begin{aligned} R(x) &= 1 + \left(\sum_{j=1}^{J} x^{m_j} \right) Z(S_4, R(x)) \\ &= 1 + \frac{1}{24} \left(\sum_{j=1}^{J} x^{m_j} \right) \Big(R(x)^4 + 6R(x)^2 R(x^2) + 3R(x^2)^2 + 8R(x)R(x^3) + 6R(x^4) \Big), \end{aligned} \tag{5}$$

since the root can bear any of the $J = |\Sigma|$ given masses $\Sigma = \{m_1, \ldots, m_J\}$. We again introduce auxiliary sequences $r_2[n]$, $r_3[n]$, and $r_4[n]$ by convolution using eq. (3). Then, the above translates to a recursion for $r[n]$, which now reads as follows:

$$\begin{aligned} r[n] = \sum_{j=1}^{J} \Bigg(& \frac{1}{24} r_4[n - m_j] + \frac{1}{4} \sum_{k=0}^{\lfloor (n-m_j)/2 \rfloor} r[k] r_2[n - m_j - 2k] + \frac{1}{8} \sum_{k=0}^{(n-m_j)/2} r[k] r[(n - m_j - 2k)/2] \\ & + \frac{1}{3} \sum_{k=0}^{\lfloor (n-m_j)/3 \rfloor} r[k] r[n - m_j - 3k] + \frac{r[(n-m_j)/4]}{4} \Bigg), \end{aligned} \tag{6}$$

where we interpret $r_i[m]$ as 0 again if the argument m is negative or not an integer.

Proposition 2. *The exact number of glycan topologies with mass M over a multiset Σ of integer masses can be computed in $O(|\Sigma|\, M^2)$ time and $O(M)$ space using recurrences* (3) *and* (6)*.*

It is not hard to tweak the algorithm a little further to get rid of the factor $|\Sigma|$, see the following section. For counting glycan topologies with n vertices over an alphabet of monosaccharides, we may assume $\Sigma = \{1, \ldots, 1\}$, and (6) simplifies to:

$$\begin{aligned} r[n] = |\Sigma| \cdot \Bigg(& \frac{1}{24} r_4[n-1] + \frac{1}{4} \sum_{k=0}^{\lfloor (n-1)/2 \rfloor} r[k] r_2[n - 1 - 2k] + \frac{1}{8} \sum_{k=0}^{(n-1)/2} r[k] r[(n - 1 - 2k)/2] \\ & + \frac{1}{3} \sum_{k=0}^{\lfloor (n-1)/3 \rfloor} r[k] r[n - 1 - 3k] + \frac{r[(n-1)/4]}{4} \Bigg) \end{aligned} \tag{7}$$

where $r_i[m] = 0$ for non-integer m. We reach:

Proposition 3. *The exact number of glycan topologies with n vertices over an alphabet Σ of monosaccharides can be computed in $O(n^2)$ time and $O(n)$ space using recurrences* (3) *and* (7)*.*

We have depicted the number of glycan topologies with a particular number of vertices n and a particular alphabet Σ in Table 1; the exact numbers are available with the Supplementary Material. We have implemented recurrence (3) and (7) in Groovy, using the BigInteger data type for arbitrary precission; the source code is also available as Supplementary Material. Computing all numbers in Table 1 required less than one second on a laptop computer.

n	$\lvert\Sigma\rvert = 1$	2	3	4	5
0	0	0	0	0	0
1	1	2	3	4	5
2	1	4	9	16	25
3	2	14	45	104	200
4	4	52	246	752	1 800
5	9	214	1 485	5 996	17 850
6	19	904	9 369	50 288	186 750
7	45	4 038	61 947	440 784	2 039 500
8	106	18 508	421 668	3 980 384	22 951 875
9	260	87 008	2 939 562	36 796 880	264 378 750
10	643	416 388	20 870 451	346 453 280	3 101 517 625
11	1 624	2 023 618	150 414 273	3 310 840 888	36 928 441 500
12	4 138	9 956 116	1 097 453 718	32 030 094 848	445 106 927 500
13	10 683	49 500 150	8 090 551 569	313 079 922 520	$5.420461 \cdot 10^{12}$
14	27 790	248 292 168	60 171 931 314	$3.087208 \cdot 10^{12}$	$6.659108 \cdot 10^{13}$
15	72 917	1 254 995 626	450 928 231 767	$3.067365 \cdot 10^{13}$	$8.242903 \cdot 10^{14}$
16	192 548	6 385 679 116	$3.401674 \cdot 10^{12}$	$3.067825 \cdot 10^{14}$	$1.027085 \cdot 10^{16}$
17	511 624	32 682 059 354	$2.581084 \cdot 10^{13}$	$3.086133 \cdot 10^{15}$	$1.287208 \cdot 10^{17}$
18	1 366 424	168 134 692 700	$1.968556 \cdot 10^{14}$	$3.120547 \cdot 10^{16}$	$1.621515 \cdot 10^{18}$
19	3 666 930	868 976 782 252	$1.508301 \cdot 10^{15}$	$3.169846 \cdot 10^{17}$	$2.052025 \cdot 10^{19}$
20	9 881 527	$4.509787 \cdot 10^{12}$	$1.160430 \cdot 10^{16}$	$3.233206 \cdot 10^{18}$	$2.607538 \cdot 10^{20}$
21	26 730 495	$2.349231 \cdot 10^{13}$	$8.961193 \cdot 10^{16}$	$3.310110 \cdot 10^{19}$	$3.325763 \cdot 10^{21}$
22	72 556 208	$1.227909 \cdot 10^{14}$	$6.943510 \cdot 10^{17}$	$3.400289 \cdot 10^{20}$	$4.256133 \cdot 10^{22}$
23	197 562 840	$6.437935 \cdot 10^{14}$	$5.396702 \cdot 10^{18}$	$3.503674 \cdot 10^{21}$	$5.463521 \cdot 10^{23}$
24	539 479 354	$3.384957 \cdot 10^{15}$	$4.206303 \cdot 10^{19}$	$3.620370 \cdot 10^{22}$	$7.033162 \cdot 10^{24}$
25	1 477 016 717	$1.784373 \cdot 10^{16}$	$3.286971 \cdot 10^{20}$	$3.750628 \cdot 10^{23}$	$9.077156 \cdot 10^{25}$
26	4 053 631 757	$9.428791 \cdot 10^{16}$	$2.574698 \cdot 10^{21}$	$3.894837 \cdot 10^{24}$	$1.174310 \cdot 10^{27}$
27	11 149 957 667	$4.993279 \cdot 10^{17}$	$2.021222 \cdot 10^{22}$	$4.053507 \cdot 10^{25}$	$1.522549 \cdot 10^{28}$
28	30 732 671 572	$2.649753 \cdot 10^{18}$	$1.589973 \cdot 10^{23}$	$4.227266 \cdot 10^{26}$	$1.978092 \cdot 10^{29}$
29	84 871 652 538	$1.408810 \cdot 10^{19}$	$1.253116 \cdot 10^{24}$	$4.416855 \cdot 10^{27}$	$2.574816 \cdot 10^{30}$
30	234 802 661 446	$7.503621 \cdot 10^{19}$	$9.893782 \cdot 10^{24}$	$4.623128 \cdot 10^{28}$	$3.357491 \cdot 10^{31}$

Table1. Number of glycan toplogies for varying number of vertices n and size of the alphabet Σ.

As an exemplary application of our method for counting glycan topologies of a particular mass, we have calculated the number of topologies for the monosaccharide residues of fucose, hexose, and N-acetylhexosamines. Assuming integer masses this results in the alphabet $\Sigma = \{146, 162, 203\}$. We also computed these numbers for the sub-alphabets $\{162, 203\}$ and $\{162\}$, see Fig. 1. Clearly, masses of monosaccharides are not integers, and the (monoisotopic) mass of a hexose residue $C_6H_{10}O_5$ is more accurately given as 162.052823 Dalton. (One Dalton, defined as $\frac{1}{12}$ of the mass of a Carbon-12 isotope, is approximately the mass of a single neutron.) Restricting ourselves to rounded integer masses is done solely for the ease of presentation. With the improved running time and space requirements of our methods, we can take into account two or three decimal places of monosaccharide masses, allowing more accurate estimates. Using a "mass blowup" or multiplier $b \in \mathbb{R}$ for all masses before rounding will, according to Theorem 1 below, result in a factor $O(b^2)$ increase in running time. Larger blowup factors are also required to avoid rounding error accumulation for large masses M; but for the mass ranges relevant in applications, this is not a severe problem, as masses of all elements are "almost integer".

We observe strong combinatorial effects: For most masses in the relevant mass range of up to 3000 Dalton we *do not find any* glycan topology. This is related to the fact that most masses cannot be decomposed over these alphabets, and that the number of decompositions varies strongly, a fact previously observed for other biomolecules (Böcker and Lipták, 2005, 2007). Even for non-zero entries, the number of topologies varies between roughly $3 \cdot 10^4$ and 10^{13} for the alphabet $\Sigma = \{146, 162, 203\}$ and masses $M = 2700, \ldots, 3000$.

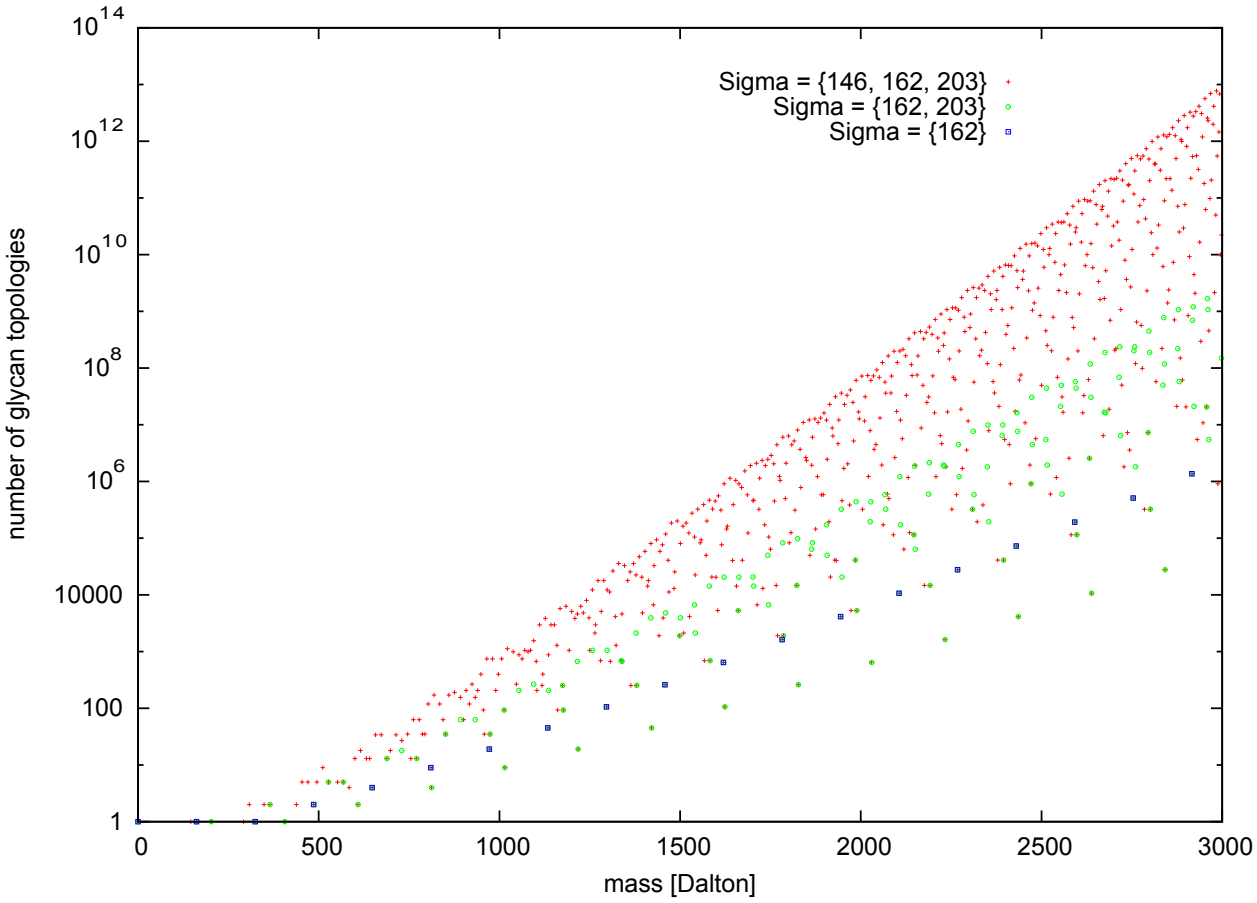

Figure1. Number of glycan topologies for varying mass M and the monosaccharide alphabets $\Sigma = \{162\}$ (hexose), $\Sigma = \{162, 203\}$ (hexose, N-acetylhexosamines), and $\Sigma = \{146, 162, 203\}$ (fucose, hexose, N-acetylhexosamines). Y-axis (number of topologies) is logarithmic. The plot is relatively sparse: Whenever no point is present in the plot, this implies that there is no glycan topology of the corresponding mass.

4 Counting d-ary trees

We now consider the most general case of counting rooted trees with maximum outdegree d. Equivalently, one can count trees whose vertices all have outdegree exactly d by adding artificial leaves. In addition, we are given a multiset $\Sigma = \{m_1, m_2, \ldots, m_J\}$ of masses again.

As in the previous sections, we can make use of the classical approach involving generating functions and Pólya's enumeration method. As we will see, the only new difficulty that arises is the number of terms in the cycle index, which increases quite rapidly with d. As above, we assume that $n = M$ is the mass we want to decompose, to allow for a uniform presentation of our calculations. Once again, let $r[n]$ be the number of rooted trees with total mass n, where every internal vertex has out-degree d (and thus every internal vertex except for the root has degree $d+1$). To simplify our presentation, we will leave out the index d throughout this section. Let

$$R(x) := \sum_{n \geq 0} r[n] x^n$$

be the associated generating function. Then $R(x)$ satisfies the functional equation

$$R(x) = 1 + \Big(\sum_{j=1}^{J} x^{m_j} \Big) Z(S_d, R(x)). \tag{8}$$

If we want to turn this equation directly into a recursion for $r[n]$, we find that the number of terms in such a recursion is essentially the number of terms in the cycle index $Z(S_d)$, which is the number of partitions of d. Since the number of partitions is asymptotically $\frac{1}{4\sqrt{3}d} e^{\pi\sqrt{2d/3}}$ and, hence, the running time would be superpolynomial in d, we have to take some additional care in order to obtain an efficient algorithm: instead of using the coefficients of the powers of R as auxiliary sequences, we consider lower-order cycle indices and make use of the identity

$$Z(S_k; s_1, s_2, \ldots, s_k) = \frac{1}{k} \sum_{j=1}^{k} s_j Z(S_{k-j}; s_1, s_2, \ldots, s_{k-j}), \tag{9}$$

see (Harary and Palmer, 1973). This suggests that we should consider the auxiliary power series

$$R^{(k)}(x) = \sum_{n \geq 0} r^{(k)}[n] x^n = Z(S_k, R(x))$$

for $k = 2, \ldots, d$ ($R^{(0)}(x) = 1$ and $R^{(1)}(x) = R(x)$ are trivial) and determine the coefficients of R and all $R^{(k)}$ according to the following steps:

1. Initialize $r^{(k)}[0] = 1$ for all $k = 0, \ldots, d$ and $r^{(0)}[n] = 0$ for all $n \geq 1$. In the following, we also assume $r^{(k)}[n] = 0$ if $n < 0$.
2. The n-th step consists of computing $r^{(k)}[n]$ for $k = 1, \ldots, d$. By our functional equation (8), we have

$$r[n] = r^{(1)}[n] = \sum_{j=1}^{J} r^{(d)}[n - m_j]. \tag{10}$$

3. Moreover, the recursion (9) yields

$$r^{(k)}[n] = \frac{1}{k} \sum_{j=1}^{k} \sum_{m=0}^{\lfloor n/j \rfloor} r[m] r^{(k-j)}[n - jm] \tag{11}$$

for $k = 2, 3, \ldots, d$.

Let us analyze the performance of this algorithm: in order to evaluate the sum in (10), we need $O(J)$ additions. For the double sum in (11), $O(n \log k)$ operations are required for every $k > 1$ (and $O(n)$ for $k = 1$), since

$$\sum_{j=1}^{k} \frac{n}{j} = nH_k = O(n \log k),$$

where H_k are the harmonic numbers. Thus we have a total of $O(dn \log d)$ operations. Altogether, this means $O(dn^2 \log d + Jn)$ operations (additions and multiplications) to compute $r[n]$. Furthermore, we have to store $O(dn)$ numbers of the form $r^{(k)}[n]$.

Theorem 1. *The exact number of rooted d-ary trees with vertices labeled with masses from a set Σ and total vertex mass M can be computed in $O(dM^2 \log d + |\Sigma|\, M)$ time and $O(dM)$ space.*

In all our considerations, we were only interested in counting the number of rooted trees. However, it is not hard to extend our ideas to the problem of counting unrooted trees. By a well-known theorem of Otter (1948), the number of edge-rooted representations of a tree is precisely one less than the number of rooted representations, except when the tree has a symmetry edge (in which case the two numbers are the same). This can be translated to the world of generating functions as follows: let $t[n]$ be the number of non-isomorphic trees whose internal vertices all have degree $d + 1$ and whose total mass is n. Denote the associated generating function by $T(x)$. Then

$$T(x) = \Big(\sum_{j=1}^{J} x^{m_j} \Big) Z(S_{d+1}, R(x)) - \frac{1}{2} \left((R(x) - 1)^2 - (R(x^2) - 1) \right). \tag{12}$$

n	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$r[0,n]$	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$r[1,n]$	1	1	1	2	4	9	19	45	106	260	643	1624	4138	10683	27790	72917
$r[2,n]$	1	1	2	3	7	15	35	81	198	485	1220	3090	7944	20580	53820	141659
$r[3,n]$	1	1	2	4	8	18	42	99	241	598	1504	3833	9876	25676	67289	177565
$r[4,n]$	1	1	2	4	9	19	45	106	260	643	1624	4138	10683	27790	72917	192548
$t[n]$	1	1	1	1	2	3	6	10	21	42	94	204	473	1098	2633	6353

Table2. Numerical values of $r[k,n]$ and $t[n]$ in the special case $d=4$, $\Sigma=\{1\}$. Note in particular that $r[1,n] = r[4,n-1]$ in this example.

The first term counts rooted trees (rooted at an internal vertex), the second one edge-rooted trees, rooted at an edge between two internal vertices, which can be regarded as pairs of rooted trees. The term $R(x^2)-1$ is a correction term for trees with a symmetry edge.

Now we can make use of (9) once again to determine first

$$r^{(d+1)}[n] = \frac{1}{d+1}\sum_{j=1}^{d+1}\sum_{m=0}^{\lfloor n/j\rfloor} r[m]r^{(d+1-j)}[n-jm]$$

for all n and then deduce from (12) that

$$\begin{aligned} t[n] &= \sum_{j=1}^{J} r^{(d+1)}[n-m_j] - \frac{1}{2}\sum_{m=0}^{n} r[m]r[n-m] + r[n] + \frac{1}{2}r[n/2] \\ &= \sum_{j=1}^{J} r^{(d+1)}[n-m_j] - r^{(2)}[n] + r[n] + r[n/2] \end{aligned}$$

for all $n>0$, where the last term is taken to be 0 for odd n. This merely requires another $O(n\log d + J)$ operations. To illustrate our algorithms, we provide numerical values of all $r[k,n]$ and $t[n]$ for $n\le 15$ in the special case $d=4$, $\Sigma=\{1\}$ in Table 2.

5 Approximating the growth

The tree families we are investigating here (with or without masses) are well known to belong to a much wider class of trees for which the number of rooted trees of given order n asymptotically grows like $A\cdot n^{-3/2}\cdot c^n$ for some constants A and c. See for example (Bell et al., 2006) for a rather general treatment of the topic. The number of unrooted trees follows a similar asymptotic formula, the only difference being an exponent 5/2 instead of 3/2. This follows from the fact that the associated generating functions have a square root singularity as their dominant singularity. In this section, we briefly outline how this singularity is obtained, see (Harary and Palmer, 1973, Section 9.5) or (Flajolet and Sedgewick, 2009, Section VII.5) for more details. Let us recall the functional equation (8):

$$R(x) = 1 + \Big(\sum_{j=1}^{J} x^{m_j}\Big) Z(S_d, R(x)) = 1 + \Big(\sum_{j=1}^{J} x^{m_j}\Big) Z(S_d; R(x), R(x^2), \ldots, R(x^d)).$$

We assume (without loss of generality) that the masses m_j have a greatest common divisor (GCD) of 1, since one could otherwise divide by the common divisor. In this case, there is a unique dominant singularity (which is necessarily real and positive) on the circle of convergence. By the implicit function theorem, this singularity ρ is obtained as the solution $(r,x)=(r_0,\rho)$ of the system of equations given by

$$F(r,x) = 1 - r + \Big(\sum_{j=1}^{J} x^{m_j}\Big) Z(S_d; r, R(x^2), \ldots, R(x^d)) = 0 \tag{13}$$

and

$$1 = \Big(\sum_{j=1}^{J} x^{m_j}\Big)\frac{\partial}{\partial r} Z(S_d, r, R(x^2), \ldots, R(x^d))$$
$$= \Big(\sum_{j=1}^{J} x^{m_j}\Big) Z(S_{d-1}; r, R(x^2), R(x^3), \ldots, R(x^{d-1})). \tag{14}$$

The coefficients of the asymptotic expansion of $R(x)$ around ρ can be expressed in terms of derivatives of the function F defined in (13):

$$R(x) \sim r_0 - \sqrt{\frac{2F_x(r_0,\rho)}{F_{rr}(r_0,\rho)}}(\rho - x)^{1/2}.$$

Now one can invoke the Flajolet-Odlyzko singularity analysis (Flajolet and Sedgewick, 2009, Chapter VI) to obtain

$$r[n] \sim A \cdot n^{-3/2}(1/\rho)^n, \tag{15}$$

where

$$A = \sqrt{\frac{\rho F_x(r_0,\rho)}{2\pi F_{rr}(r_0,\rho)}},$$
$$F_x(r_0,\rho) = \Big(\sum_{j=1}^{J} m_j \rho^{m_j-1}\Big) Z(S_d; r_0, R(\rho^2), R(\rho^3), \ldots)$$
$$+ \Big(\sum_{j=1}^{J} \rho^{m_j}\Big) \sum_{\ell=2}^{d} \rho^{\ell-1} R'(\rho^\ell) Z(S_{d-\ell}; r_0, R(\rho^2), R(\rho^3), \ldots),$$
$$F_{rr}(r_0,\rho) = \Big(\sum_{j=1}^{J} \rho^{m_j}\Big) Z(S_{d-2}; r_0, R(\rho^2), R(\rho^3), \ldots)).$$

In principle, one can even extend this approximation to a full asymptotic expansion, and we can also deduce the asymptotic behavior for unrooted trees. The generating function $T(x)$ inherits the location of the dominant singularity ρ, the only difference is the type of singularity. One obtains, with

$$B = \sqrt{\frac{\rho F_x(r_0,\rho)}{2\pi F_{rr}(r_0,\rho)}}\Bigg(\Big(\sum_{j=1}^{J} m_j \rho^{m_j}\Big) Z(S_d, r_0, R(\rho^2), R(\rho^3), \ldots)$$
$$+ \Big(\sum_{j=1}^{J} \rho^{m_j}\Big) \sum_{k=2}^{d} \rho^{k} R'(\rho^k) Z(S_{d-k}, r_0, R(\rho^2), R(\rho^3), \ldots)\Bigg),$$

that the number of unrooted trees $t[n]$ (defined as in the previous section) satisfies

$$t[n] \sim B \cdot n^{-5/2}(1/\rho)^n. \tag{16}$$

The only practical task remaining is to compute the constants A, B, and $1/\rho$, a problem that we briefly consider in the following section.

6 Computing the approximation constants

While it is comforting to know that asymptotic formulas of the form (15) and (16) hold, one would obviously also like to determine numerical values for the constants that occur in these formulas. We will see below that this is not sensible for counting glycans of a particular mass. But there exist experimental techniques for the analysis of glycans that cannot measure the mass of the glycan but only its size n, one prominent example

d	A	B	$1/\rho$
2	0.7916031836	1.255108880	2.483253536
3	0.5178759065	0.6563186958	2.815460033
4	0.4621373461	0.5626104567	2.911037772
5	0.4464847137	0.5407688237	2.941025361
6	0.4418238966	0.5357072863	2.950834412
7	0.4404523304	0.5347787892	2.954104926
8	0.4400624723	0.5347433285	2.955204633
9	0.4399569574	0.5348289013	2.955575751
10	0.4399304158	0.5348914403	2.955701181

Table3. Approximation constants $A, 1/\rho$ for the number of rooted trees with n vertices using eq. (15), and constants $B, 1/\rho$ for the number of unrooted trees using eq. (16). Vertices have maximum degree d and are not labeled.

$\lvert\Sigma\rvert$	A	$1/\rho$
1	0.4621373461	2.911037772
2	0.4359963877	5.603311188
3	0.4286144321	8.305741452
4	0.4251715830	11.01087067
5	0.4231864914	13.71711627

$\lvert\Sigma\rvert$	A	$1/\rho$
6	0.4218970932	16.42393156
7	0.4209928959	19.13107690
8	0.4203239993	21.83843057
9	0.4198092582	24.54592417
10	0.4194009528	27.25351630

Table4. Approximation constants $A, 1/\rho$ for the number of glycan topologies with n vertices (monosaccharides) using eq. (15). Vertices are labeled from the alphabet Σ.

being gel electrophoresis (Jayo et al., 2012). For such techniques, it can again be important to determine the number of possible glycan structures as well as to estimate the exponential growth. Obviously, we can determine the exact number using the techniques presented above. But as we will see, we can approximate the true number using eqs. (15) and (16) with a precission that, even for moderate size n, will be sufficient for many applications. In addition, estimating the asymptotic growth allows us to get a rough idea of the number of structures we have to expect for a particular alphabet Σ and size n.

Computing the constants for (15) and (16) can be done by means of a simple recipe: note that the radius of convergence of $R(x^k)$ is $\rho^{1/k}$. It follows that the series for $R(x^k)$ converges exponentially around $x = \rho$ for $k \geq 2$, and one can approximate $R(x^k)$ very well by a finite number of terms in its Taylor expansion. Hence one can apply the following technique to determine ρ and r_0:

- Determine a sufficiently large number of terms of the sequence $r[n]$, say one hundred terms.
- Replace $R(x^k)$ by $\sum_{n=0}^{100} r[n]x^{kn}$ in the two equations (13) and (14).
- Solve the resulting system of equations numerically for the two unknowns x and r to obtain numerical values for ρ and r_0 respectively.
- The constants A and B that occur in the two asymptotic formulas (15) and (16) are obtained by plugging in ρ and r_0 and replacing $R(x^k)$ by a finite sum for all k again.

Of course, the approximations become more accurate if more terms of the sequence $r[n]$ are determined, but typically very few terms are sufficient for an excellent accuracy in view of the exponential convergence.

Let us illustrate this recipe by some examples: In Table 3 we give constants $A, B, 1/\rho$ for rooted and unrooted d-ary trees with $|\Sigma| = 1$ and $d = 2, \ldots, 10$. In Table 4 we focus on the case $d = 4$ particularly interesting for glycans, and list constants for rooted trees and $|\Sigma| = 1, \ldots, 10$. Note that a similar approximation is possible for glycans of a particular mass (see Sec. 3) but this is not useful in practice: Due to the combinatorial nature of decomposing integers, we can expect this approximation to be completely uninformative for any relevant mass range of up to 100 000 Dalton.

The approximations using Tables 3 and 4 are very accurate even for relatively small n: For example, for $n = 20$ and $|\Sigma| = 5$ we approximate $r[n] \approx 2.6317 \cdot 10^{20}$ using eq. (15) and Table 4, whereas the true number is $2.6075 \cdot 10^{20}$, compare to Table 1. Even for this moderate n, the relative error is below 1 %.

7 Ordered trees and glycan structures

If we consider glycan structures instead of glycan topologies, then we can assume that the four out-links of a monosaccharide have a specific order. This means that we are counting *ordered trees* rather than unordered ones, where the order of the children of a vertex matters. (Recall that we may assume that we are counting trees with n interior vertices such that all interior vertices have maximum degree. To this end, the order of children of a vertex is equivalent to an index.) The enumeration of ordered trees is somewhat simpler, and no cycle indices are needed. If we denote the number of ordered 4-ary trees by $r^*[n]$, then the associated generating function $R^*(x)$ satisfies

$$R^*(x) = 1 + xR^*(x)^4,$$

and it is well known (and can be deduced from Lagrange's inversion formula, amongst others) that

$$r^*[n] = \frac{(4n)!}{n!(3n+1)!} = \frac{1}{3n+1}\binom{4n}{n},$$

a generalized Catalan number. For d-ary trees, simply replace "4" and "3" in the above formula by "d" and "$d-1$", respectively. If the vertices receive labels from an alphabet Σ, then we only have to multiply by a factor $|\Sigma|^n$, and it is also easy to obtain an asymptotic approximation from Stirling's formula.

Things become more interesting if we also consider the situation where each vertex is assigned a mass from a multiset Σ and we are counting trees by total mass rather than number of vertices. In this case, the functional equation for the generating function reads

$$R^*(x) = 1 + \left(\sum_{j=1}^{J} x^{m_j}\right) R^*(x)^d.$$

Now an explicit formula for the number of trees with total mass M is generally no longer available, but the functional equation can be turned into an efficient algorithm for computing the coefficients by a similar approach as before. Indeed, instead of $O(dn^2 \log d + |\Sigma|\, n)$ time and $O(dn)$ space, $O(n^2 \log d + |\Sigma|\, n)$ time and $O(n \log d)$ space are sufficient: in order to determine the coefficients of the d-th power $R^*(x)^d$, one can use the usual method of exponentiation by squaring, so that only $O(\log d)$ auxiliary functions are needed.

The asymptotic behavior is similar to unordered trees: again, one has $r^*[n] \sim A \cdot n^{-3/2} \cdot \rho^{-n}$, where ρ is now a solution of the equation

$$\sum_{j=1}^{J} \rho^{m_j} = \frac{(d-1)^{d-1}}{d^d},$$

and

$$A = \sqrt{\frac{d\sum_{j=1}^{J} j\rho^{m_j}}{2\pi(d-1)^3 \sum_{j=1}^{J} \rho^{m_j}}}.$$

8 Conclusion

We have presented a simple recipe for counting d-ary tree structures with a given number of vertices or total mass both approximately (by determining the asymptotic growth constants) and by an exact algorithm whose running time and space requirement improves on previously suggested algorithms. Our improvements are particularly relevant for determining the number of glycan topologies of a particular mass, using modern mass spectrometry instruments with high mass accuracy. Our approximation results allow us to give accurate estimates of the number of glycan topologies with n monosaccharides, without having to solve the recurrences. Our approximation results show that the number of glycan topologies with n vertices growths practically as fast as the number of peptides with n amino acids, if the monosaccharide alphabet Σ reaches $|\Sigma| \geq 7$. Glycans are certainly not the only chemically and biologically relevant example to which our methods apply, as we have noted for the example of alkanes. So, we believe that this paper will provide a useful reference for related counting problems in the future.

Acknowledgments. We thank Birte Kehr for preparing Figure 1. This material is based upon work supported financially by the National Research Foundation of South Africa under grant number 70560.

Bibliography

Bell, J. P., Burris, S. N., and Yeats, K. A. (2006). Counting rooted trees: the universal law $t(n) \sim C\rho^{-n}n^{-3/2}$. *Electron. J. Combin.*, 13(1):Research Paper 63, 64 pp. (electronic).

Böcker, S., Kehr, B., and Rasche, F. (2011). Determination of glycan structure from tandem mass spectra. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 8(4):976–986.

Böcker, S. and Lipták, Zs. (2005). Efficient mass decomposition. In *Proc. of ACM Symposium on Applied Computing (ACM SAC 2005)*, pages 151–157, Santa Fe, USA. ACM Press.

Böcker, S. and Lipták, Zs. (2007). A fast and simple algorithm for the Money Changing Problem. *Algorithmica*, 48(4):413–432.

Cayley, A. (1881). On the analytical forms called trees. *American Journal of Mathematics*, 4:266–268.

Ethier, M., Saba, J. A., Spearman, M., Krokhin, O., Butler, M., Ens, W., Standing, K. G., and Perreault, H. (2003). Application of the StrOligo algorithm for the automated structure assignment of complex N-linked glycans from glycoproteins using tandem mass spectrometry. *Rapid Commun. Mass Spectrom.*, 17(24):2713–2720.

Flajolet, P. and Sedgewick, R. (2009). *Analytic Combinatorics*. Cambridge University Press. Freely available from `http://algo.inria.fr/flajolet/Publications/book.pdf`.

Gaucher, S. P., Morrow, J., and Leary, J. A. (2000). STAT: a saccharide topology analysis tool used in combination with tandem mass spectrometry. *Anal. Chem.*, 72(11):2331–2336.

Goldberg, D., Bern, M. W., Li, B., and Lebrilla, C. B. (2006). Automatic determination of O-glycan structure from fragmentation spectra. *J. Proteome Res.*, 5(6):1429–1434.

Harary, F. and Palmer, E. M. (1973). *Graphical enumeration*. Academic Press, New York.

Harary, F., Robinson, R. W., and Schwenk, A. J. (1975). Twenty-step algorithm for determining the asymptotic number of trees of various species. *Journal of the Australian Mathematical Society Series A*, 20(4):483–503.

Henze, H. R. and Blair, C. M. (1931). The number of structurally isomeric alcohols of the methanol series. *J. Am. Chem. Soc.*, 53(8):3042–3046.

Jayo, R. G., Li, J., and Chen, D.D. (2013). Capillary electrophoresis mass spectrometry for the characterization of O-acetylated N-glycans from fish serum. *Anal. Chem.*, 84(20):8756–8762.

Otter, R. (1948). The number of trees. *The Annals of Mathematics*, 49(3):583–599.

Palmisano, G., Antonacci, D., and Larsen, M. R. (2010). Glycoproteomic profile in wine: a 'sweet' molecular renaissance. *J. Proteome Res.*, 9(12):6148–6159.

Pólya, G. (1937). Kombinatorische Anzahlbestimmungen für Gruppen, Graphen und chemische Verbindungen. *Acta Mathematica*, 68(1):145–254.

Rains, E. M. and Sloane, N. J. A. (1999). On Cayley's enumeration of alkanes (or 4-valent trees). *Journal of Integer Sequences*, 2:(electronic).

Raman, R., Raguram, S., Venkataraman, G., Paulson, J. C., and Sasisekharan, R. (2005). Glycomics: an integrated systems approach to structure-function relationships of glycans. *Nat. Methods*, 2(11):817–824.

Trinajstić, N., Jeričević, Ž., Knop, J. V., Müller, W. R., and Szymanski, K. (1983). Computer generation of isomeric structures. *Pure and Applied Chemistry*, 55(2):379–390.

Varki, A., Cummings, R. D., Esko, J. D., Freeze, H. H., Stanley, P., Bertozzi, C. R., Hart, G. W., and Etzler, M. E., editors (2009). *Essentials of Glycobiology*. Cold Spring Harbor Laboratory Press, second edition. Freely available from `http://www.ncbi.nlm.nih.gov/books/NBK1908/`.