

Asymptotic isomer enumeration in chemistry

John Knopfmacher

*Centre for Applied Analysis and Number Theory, c/o Department of Mathematics,
University of the Witwatersrand, P.O. Wits 2050, Johannesburg, South Africa
Currently also: c/o Department of Mathematics, University of Melbourne, Parkville,
Victoria 3052, Australia*

Received 9 October 1997; revised 22 April 1998

In the theory of chemical isomerism, considerable attention has been paid in the past to the enumeration of all isomers (of a chosen type) of all *molecules* with a particular molecular formula of interest, having an appropriate number of atoms of some special interest (e.g., carbon atoms). Some corresponding methods and results are here developed for enumerating *ensembles* or *mixtures* of molecules. The main emphasis is on *asymptotic* enumeration results.

1. Introduction

An important aspect of the theory of chemical isomerism (cf. Rouvray [8] – especially section 9, and Slanina [9] – especially sections 5.3.6 and 5.6), going back more than a century, has been the development of methods for enumerating the isomers (of varying kinds) of molecules having molecular formulae of some particular significance, e.g., the formulae C_nH_{2n+2} for normal, unbranched hydrocarbons (alkanes, paraffins). Standard texts (e.g., Slanina [9, chapter 5] and Trinajstić [10, chapter 11]) often emphasize the rapid growth with n of the number of constitutional (structural) or stereo isomers of molecules with such formulae as C_nH_{2n+2} , etc. This phenomenon for various important types of molecules led to interest in both the algebraic and the asymptotic behaviour of the number of isomers of molecules with a particular type of molecular formula involving increasing numbers n of carbon atoms.

Major contributions towards answering questions of the preceding type were published by G. Pólya in 1937, in a famous *Acta Mathematica* (Sweden) paper which was reprinted in translation [7] in 1987, together with an up-dating commentary by R.C. Read. For example, let ρ_n denote the total number of constitutional (structural) isomers of alkane molecules (paraffins) of type C_nH_{2n+2} , and let R_n denote the total number of constitutional isomers of alkanol (alcohol) molecules of type $C_nH_{2n+1}OH$. Pólya [7] established that

$$\rho_n \sim C_1 q_1^n n^{-5/2}, \quad R_n \sim C_2 q_1^n n^{-3/2} \quad \text{as } n \rightarrow \infty,$$

where C_1 , C_2 , q_1 are explicitly defined positive constants, with $q_1 = 2.85\dots$, and $A_n \sim B_n$ as $n \rightarrow \infty$ means that $\lim_{n \rightarrow \infty} A_n/B_n = 1$. He also gave various similar types of asymptotic formulae as $n \rightarrow \infty$ for corresponding isomer counting numbers of certain other types of molecules with n carbon atoms.

If one takes a general viewpoint of theoretical constitutional or stereo chemistry, as propounded, e.g., by Dugundji and Ugi [1], Dugundji et al. [2], or Ugi et al. [11], the above remarkable conclusions of Pólya suggest some natural *further* asymptotic isomer enumeration problems as follows, whose solutions require *additional* techniques to those of Pólya [7]:

Consider an *ensemble* of molecules $EM(A)$ as in Dugundji and Ugi [1], or Dugundji et al. [2], where A is some set of atoms. (By this is meant any compound or collection of chemical species which can be formed from A using each atom in A exactly once; here the word “ensemble” is *not* used in a thermodynamic or statistical mechanics sense. The set of all the $EM(A)$ is called the *family of isomeric ensembles of molecules* of A in [1,2], and denoted by $FIEM(A)$.) Next, suppose that Γ is a set of molecular formulae of interest, e.g., the formulae C_nH_{2n+2} for alkanes ($n = 1, 2, 3, \dots$). Instead of concentrating only on ensembles $EM(A)$ and $FIEM(A)$ as defined in [1,2], we shall *drop* the condition that all atoms in A appear exactly once in each ensemble $EM(A)$, and *consider instead* the family $FIEM_\Gamma(A)$ of all isomeric ensembles whose molecules have molecular formulae lying in Γ , and their atoms all lie in A but may be repeated. One might perhaps call any isomeric ensemble E in $FIEM_\Gamma(A)$ a Γ -*mixture*. For example, if Γ is the set of formulae C_nH_{2n+2} ($n = 1, 2, 3, \dots$), one might call E a *mixed* alkane (alkane-mixture), or a *mixed* paraffin (paraffin-mixture).

Now suppose that Γ consists of various formulae which all involve one or more atoms of a certain single type of interest, e.g., carbon atoms. In the earlier Pólya-type theory one investigates the number $\sigma_n^*(\gamma)$ of all isomers of *molecules* having some single formula γ in Γ (where γ involves n atoms of the special kind selected). However, it would also seem natural to consider the extension of isomerism to *ensembles* of molecules as in [1,2], and to then seek information about the number $\sigma_n(\Gamma)$ of all isomeric *ensembles* in $FIEM_\Gamma(A)$, i.e., not necessarily only single molecules, which involve exactly n of the special atoms *in total*. For example, it would be natural to investigate the total number σ_n of mixed alkanes (mixed paraffins) containing exactly n carbon atoms, counting over all their molecules *in total*.

Main aim: The main aim of the present article is to asymptotically answer some of the above-proposed enumeration problems about general chemical mixtures, for a few selected, well known sets Γ of important molecular formulae in organic chemistry. The later formulation of an appropriate *mathematical context*, and the indication of certain *general methods*, for deriving such answers may possibly, however, be as worthy of attention as the individual results to be listed explicitly.

Notes: Throughout the general discussion the term *isomer* is assumed to be taken in some fixed sense, such as either constitutional (structural), or stereo, or other type of interest. It is important to note that while they are *theoretically* feasible, for steric

reasons, many of the mathematically allowed isomers would not actually be chemically viable in a *laboratory* sense.¹

I am grateful to an anonymous reader for some valuable comments about a first draft of this paper, and also to Prof. J.C.A. Boeyens for helpful discussion of those comments.

2. Individual enumeration results

Consider the following particular sets of molecular formulae:

- (i) $\Gamma_1 = \{C_nH_{2n+2}: n = 1, 2, 3, \dots\}$, for alkanes,
- (ii) $\Gamma_2 = \{C_nH_{2n+1}X: n = 1, 2, 3, \dots\}$, for 1-fold substituted alkanes (e.g., alkanols if $X = OH$),
- (iii) $\Gamma_3 = \{C_{6+n}H_{6+2n}: n = 0, 1, 2, 3, \dots\}$, for alkyl substituted benzenes.

Along earlier lines, let $\sigma_n(\Gamma)$ denote the number of constitutional (structural) isomers of general Γ -mixtures containing n carbon atoms in total, and let $\Sigma_n(\Gamma)$ denote the corresponding number of stereoisomers, where Γ is any fixed set of molecular formulae which all involve one or more carbon atoms. Also, let $\sigma_n^* = \sigma_n^*(\gamma)$ and $\Sigma_n^* = \Sigma_n^*(\gamma)$ denote the corresponding numbers of isomers of *molecules* of type γ , for a formula γ in Γ involving exactly n carbon atoms.

By using asymptotic results of Pólya [7] regarding certain specific enumeration numbers σ_n^* , Σ_n^* as $n \rightarrow \infty$, and an independent mathematical theorem of Knopfmacher and Knopfmacher [6], the following asymptotic enumeration results for corresponding *general* Γ -mixtures with n carbon atoms *in total* will be derived in section 5 below. In these results, the symbols K_i , q_j denote certain particular positive constants, with $q_j > 1$ ($i = 1, \dots, 5$; $j = 1, 2$):

$$\sigma_n(\Gamma_1) \sim K_1 q_1^n n^{-5/2} \quad \text{as } n \rightarrow \infty, \quad (2.1)$$

$$\sigma_n(\Gamma_2) \sim K_2 q_1^n n^{-3/2} \quad \text{as } n \rightarrow \infty, \quad (2.2)$$

$$\Sigma_n(\Gamma_1) \sim K_3 q_2^n n^{-5/2} \quad \text{as } n \rightarrow \infty, \quad (2.3)$$

$$\Sigma_n(\Gamma_2) \sim K_4 q_2^n n^{-3/2} \quad \text{as } n \rightarrow \infty, \quad (2.4)$$

$$\sigma_n(\Gamma_3) \sim K_5 q_1^n n^{-3/2} \quad \text{as } n \rightarrow \infty. \quad (2.5)$$

3. A mathematical context for the results

In order to derive asymptotic enumeration estimates like those of (2.1)–(2.5) above, we shall begin by fitting the general discussion of families $FIEM_\Gamma(A)$ of Γ -mixtures within a convenient mathematical framework.

¹ Cf., say, D.J. Klein and W.A. Seitz, in: *Chemical Applications of Topology and Graph Theory*, ed. R.B. King (Elsevier, Amsterdam, 1983) pp. 430–445.

Suppose as before that all formulae in Γ involve one or more atoms of a certain single type of interest, to be called **-atoms* for the sake of reference here (e.g., carbon atoms). Suppose that Γ is the union of finite sets $\Gamma^{(1)}, \Gamma^{(2)}, \Gamma^{(3)}, \dots$, such that each formula γ in $\Gamma^{(n)}$ involves n **-atoms*, as well as other atoms perhaps. In all, let $A = A(\Gamma)$ denote the set of all types of atoms involved in one or more formulae of Γ . (In practice, A would usually be a finite set of atoms which all occur in every formula γ of Γ , but with possibly varying multiplicities as γ varies.)

Next, suppose that some type of isomerism has been fixed (e.g., constitutional, or stereo, etc.) and extended to ensembles of molecules of the type under consideration. For each formula γ in Γ , let $P(\gamma)$ denote the set of all isomers of molecules with molecular formula γ , and let $P_\Gamma = \bigcup_{\gamma \in \Gamma} P(\gamma)$ denote the union of all these sets $P(\gamma)$. Then let $G'_\Gamma = FIEM_\Gamma(A)$ denote the family of all isomeric ensembles of molecules which have molecular formulae in Γ , where $A = A(\Gamma)$ as above.

Isomerism is an equivalence relation between molecules, or ensembles of molecules, and every such ensemble E can be regarded as an unordered “string” of distinct molecules M_1, \dots, M_r (say), which may be written symbolically as

$$E = M_1 + \dots + M_r.$$

The corresponding isomer equivalence class \overline{E} of E is similarly an unordered string of the isomer equivalence classes $\overline{M}_1, \dots, \overline{M}_r$, and one may similarly write formally

$$\overline{E} = \overline{M}_1 + \dots + \overline{M}_r.$$

In such a situation, some of the molecules M_i, M_j might perhaps be isomers, so that $\overline{M}_i = \overline{M}_j$. If this possibility is taken into account, then eventually one obtains a symbolic sum

$$\overline{E} = k_1 \overline{M}'_1 + \dots + k_s \overline{M}'_s,$$

in which the molecules M'_1, \dots, M'_s are mutually non-isomeric, and $k\overline{M}$ denotes a string of k copies of \overline{M} if $k = 1, 2, 3, \dots$. This symbolic expression for \overline{E} as a “sum” is then unique apart from rearrangement of the terms $k_i \overline{M}'_i$.

The above type of symbolic decomposition of the families \overline{E} in G'_Γ into symbolic sums, suggests the following *addition* operation $+$ on G'_Γ , which corresponds naturally to the idea of combining or mixing separate Γ -mixtures. Any ensembles E_1, E_2 can, with convenient choice of notation, be written as strings of distinct molecules $\{M_1, \dots, M_t\}, \{M_{t+1}, M_{t+2}, \dots, M_r\}$ with

$$E_1 = M_1 + \dots + M_t, \quad E_2 = M_{t+1} + \dots + M_r.$$

The combined or mixed ensemble defined by joining these ensembles may then be written symbolically as

$$E_1 + E_2 = M_1 + \dots + M_t + M_{t+1} + \dots + M_r,$$

and one may similarly define

$$\overline{E}_1 + \overline{E}_2 = \overline{M}_1 + \cdots + \overline{M}_t + \overline{M}_{t+1} + \cdots + \overline{M}_r.$$

(It is, of course, assumed that Γ is such that such combinations or mixtures are *stable*, and do not give rise to totally new molecules or strings of molecules.)

In terms of standard algebraic concepts, the above remarks show that under the indicated operation of addition G_Γ forms a *commutative semigroup* with *unique additive decomposition* into the “basis” set P_Γ of all isomers of actual *molecules* with molecular formulae in Γ . P_Γ is then unlike an ordinary basis of linear algebra, in that only linear combinations with *positive* integer coefficients occur relative to P_Γ in G'_Γ . It would sometimes be algebraically convenient to allow general non-negative integer coefficients by enlarging G'_Γ to $G_\Gamma = G'_\Gamma \cup \{\theta\}$, where θ denotes a single new and *purely formal* isomeric “ensemble” involving *no* molecules.

Since all ensembles of molecules with formulae in Γ contain one or more $*$ -atoms by hypothesis, the total number $\partial(E)$ or $\partial(\overline{E})$ of $*$ -atoms involved in an ensemble E provides a natural “size” or *degree* function ∂ on G_Γ such that

- (i) $\partial(x) = 0$ if and only if $x = \theta$ in G_Γ ,
- (ii) $\partial(x + y) = \partial(x) + \partial(y)$,
- (iii) the total number $G_\Gamma^\#(n)$ of elements of degree n in G_Γ is finite for each $n = 0, 1, 2, \dots$

The *present importance* of the above mathematical considerations lies mainly in the fact that (for varying sets $\Gamma = \Gamma_1, \Gamma_2, \Gamma_3$, and varying types of isomerism) the above-stated enumeration formulae (2.1)–(2.5) are asymptotic statements as $n \rightarrow \infty$ about the number $G_\Gamma^\#(n)$ of isomeric ensembles (Γ -mixtures) in G_Γ with exactly n $*$ -atoms in total, i.e., the number of elements of degree n in the semigroup G_Γ . The main reason for introducing the formal addition operation $+$ on G_Γ is that it allows the problem of *proving* (2.1)–(2.5) to be reduced as below to questions about certain types of mathematical objects which have been studied elsewhere for quite *different* reasons. Mathematical theorems about such objects can be applied to chemical settings of the above kind whenever suitable results of the types of Pólya [7] for certain kinds of isomers of molecules have already been derived separately.

4. Applications of abstract number theory

The chemical semigroups G_Γ , for varying sets Γ of molecular formulae, and varying types of isomerism, provide some natural scientific models of mathematical objects which have been investigated for various different reasons as differing kinds of abstractions of the number theory of positive integers and their prime decompositions (cf. Knopfmacher [4,5], say). In order to see how the chemical enumeration problems may be subsumed under the appropriate mathematical theory, we therefore recall the

relevant abstract number-theoretical concepts [4,5]; we confine attention to integer-valued “size” functions, since these are sufficient for present applications:

By definition, an *arithmetical semigroup* (a.s.) is a commutative semigroup G with an identity element 1 (relative to a multiplication operation \cdot), which contains a subset P such that every element $a \neq 1$ admits unique factorization into a finite product of powers of elements of P , and which admits a positive integer-valued “norm” function $||$ on G such that

- (i) $|a| = 1$ if and only if $a = 1$ in G ,
- (ii) $|a \cdot b| = |a| \cdot |b|$, and
- (iii) the total number $G(n)$ of elements a in G with $|a| = n$ is finite for each $n = 1, 2, 3, \dots$

The elements p of P are called the (abstract) *primes* of G . If, instead of a norm function $||$, there exists a non-negative integer-valued *degree* function ∂ on G with the following properties (i')–(iii'), then G is called an *additive a.s.*:

- (i') $\partial(a) = 0$ if and only if $a = 1$ in G ,
- (ii') $\partial(a \cdot b) = \partial(a) + \partial(b)$,
- (iii') the total number $G^\#(n)$ of elements a in G with $\partial(a) = n$ is finite for each $n = 0, 1, 2, \dots$

Note that, if G is any additive a.s., then G automatically becomes an a.s. in the first sense if one chooses any fixed integer $k > 1$ and then defines $|| = ||_k$ by $|a| = k^{\partial(a)}$. However, the converse statement may not be true: e.g., consider the proto-type \mathbb{N} of all arithmetical semigroups, for which $\mathbb{N} = \{1, 2, 3, \dots\}$ and $|n| = n$.

Obviously, if G is any a.s., then the total number $P(n)$ of prime elements p in P with $|p| = n$ is finite for each $n = 2, 3, 4, \dots$. Similarly, if G is any additive a.s., then the total number $P^\#(n)$ of prime elements p in P with $\partial(p) = n$ is finite for each $n = 1, 2, 3, \dots$

Theorems which establish information about the numbers $P(n)$, or $P^\#(n)$, on the basis of hypotheses about the numbers $G(n)$, or $G^\#(n)$, are called *abstract prime number theorems*, while theorems which lead in the converse direction may be called *inverse abstract prime number theorems*. For example, the famous classical prime number theorem for \mathbb{N} above asserts that

$$\pi(x) \sim \frac{x}{\log x} \quad \text{as } x \rightarrow \infty,$$

where $\pi(x)$ is the number of ordinary primes p in \mathbb{N} with $p \leq x$; here $\mathbb{N}(n) = 1$ for $n = 1, 2, 3, \dots$ is so trivial a function that its use is only implicit.

A variety of further theorems of these kinds have been established in different contexts. Some examples of such theorems are discussed, e.g., in [4,5], but a different one of Knopfmacher and Knopfmacher [6] will be quoted and used later in order to derive the stated results (2.1)–(2.5) above. However, before doing this, we return to the

relevance of these abstract number-theoretical questions to the chemical semigroups G_Γ of Γ -mixtures:

Firstly, although the natural addition operation $+$ on G_Γ has useful properties as outlined earlier, it is more convenient to the derivation of (2.1)–(2.5) to now *formally re-designate* addition as “multiplication”. Thus for elements x, y of G_Γ , we formally re-write $x + y$ as $x \cdot y$. Then kx , for x in G and $k = 1, 2, 3, \dots$, now becomes x^k , and we shall re-denote θ by the symbol 1. After doing this, we may then note that the discussion of G_Γ in section 3 implicitly shows that G_Γ is an additive a.s. Further, the earlier set P_Γ may now be regarded as the set of all abstract primes of G_Γ . In addition, the above-stated formulae (2.1)–(2.5) may be interpreted as assertions about certain particular counting numbers $G_\Gamma^\#(n)$ as $n \rightarrow \infty$. (*Note*: Since the difference between $+$ and \cdot on G_Γ is only *formal*, we could have avoided $+$ altogether. However, the initial use of $+$ might perhaps seem more *intuitive*.)

At this stage, we shall quote some asymptotic isomer enumeration results of Pólya [7] and re-interpret them as assertions about particular counting numbers $P_\Gamma^\#(n)$ as $n \rightarrow \infty$. It is convenient to do this in the form of a table, on the basis of conclusions listed in the introduction of Pólya [7], and the commentary by Read in [7]. (The survey by Harary et al. [3] may also be consulted, but uses slightly different notation.) The type of isomerism involved for the particular semigroup G_Γ under consideration, will be clear from the discussion in section 2 above; when stereoisomers are involved the *subscript* s will here be added to $P_\Gamma^\#(n)$. In the stated results, the symbols C_i , q_j denote certain particular positive constants ($i = 1, \dots, 5$; $j = 1, 2$), with $q_1 = 2.85\dots$, $q_2 = 3.33\dots$

5. An abstract prime number theorem

It is easy to see from table 1 that all the individual chemical semigroups G_Γ chosen for scrutiny in section 4 have the property hypothesised in the following “axiom”, which may or may not be valid for general additive arithmetical semigroups:

Axiom Φ . G is an additive a.s. with the property that there exist positive constants C , q , α depending on G , with $q > 1$ and $\alpha > 1$, such that

$$P^\#(n) \sim Cq^n n^{-\alpha} \quad \text{as } n \rightarrow \infty.$$

Table 1

| | Section 2 notation | Pólya's notation | Arith. semigroup notation | Asymptotic behaviour as $n \rightarrow \infty$ |
|-------|------------------------|---------------------|------------------------------|---|
| (4.1) | $\sigma_n^*(\Gamma_1)$ | ρ_n | $P_{\Gamma_1}^\#(n)$ | $C_1 q_1^n n^{-5/2}$ |
| (4.2) | $\sigma_n^*(\Gamma_2)$ | R_n | $P_{\Gamma_2}^\#(n)$ | $C_2 q_1^n n^{-3/2}$ |
| (4.3) | $\Sigma_n^*(\Gamma_1)$ | σ_n | $P_{\Gamma_{1,s}}^\#(n)$ | $C_3 q_2^n n^{-5/2}$ |
| (4.4) | $\Sigma_n^*(\Gamma_2)$ | S_n | $P_{\Gamma_{2,s}}^\#(n)$ | $C_4 q_2^n n^{-3/2}$ |
| (4.5) | $\sigma_n^*(\Gamma_3)$ | – | $P_{\Gamma_3}^\#(n)$ | $C_5 q_1^n n^{-3/2}$ |

A simple way to see that axiom Φ is not necessarily always valid, is to consider any fixed prime element p_0 in any additive a.s. G and let $G_0 = \{1, p_0, p_0^2, p_0^3, \dots\}$. Then G_0 is also an additive a.s., but has only *one* prime element p_0 . On the other hand, it can be shown that axiom Φ is also valid for various concrete mathematical examples of arithmetical semigroups, which are suggested by geometry and graph theory and have no obvious connections with chemistry. Such semigroups are investigated by Knopfmacher and Knopfmacher [6], where the following (inverse) abstract prime number theorem is proved:

Theorem 5.1. Suppose that G is any additive a.s. satisfying axiom Φ above. Then

$$G^\#(n) \sim Kq^n n^{-\alpha} \quad \text{as } n \rightarrow \infty,$$

where

$$K = CZ_G(q^{-1}) \quad \text{and} \quad Z_G(y) = \sum_{r=0}^{\infty} G^\#(r)y^r.$$

This is not the place to consider the mathematical proof of theorem 5.1. However, it should now be clear that the earlier-stated *general* isomer enumeration results (2.1)–(2.5) for *mixtures* are all corollaries of theorem 5.1, after applications have been made of Pólya's classical *molecular* isomer enumeration results, as listed in (4.1)–(4.5) of table 1.

6. Final remarks

Different mathematical settings (cf. [4,5], say), as well as some different chemical enumeration results of Pólya [7] show that axiom Φ above does *not* embrace all possible important types of asymptotic behaviour which naturally occurring arithmetical semigroups can satisfy. In particular, some of Pólya's other chemical enumeration results suggest the consideration of one or more further "axioms" in place of axiom Φ . The mathematical treatment of such other cases, which involve *different* asymptotic formulae, requires *different* techniques from those used to prove theorem 5.1. It is hoped to pursue such investigations and applications in some subsequent papers.

References

- [1] J. Dugundji and I. Ugi, An algebraic model of constitutional chemistry as a basis for chemical computer programs, *Topics Current Chem.* 39 (1973) 19–64.
- [2] J. Dugundji, P. Gillespie, D. Marquarding and I. Ugi, Metric spaces and graphs representing the logical structure of chemistry, in: *Chemical Applications of Graph Theory*, ed. A.T. Balaban (Academic Press, 1976) pp. 107–174.
- [3] F. Harary, E.M. Palmer, R.W. Robinson and R.C. Read, Pólya's contributions to chemical enumeration, in: *Chemical Applications of Graph Theory*, ed. A.T. Balaban (Academic Press, 1976) pp. 11–24.

- [4] J. Knopfmacher, *Abstract Analytic Number Theory* (North-Holland, 1975; Dover, 1990).
- [5] J. Knopfmacher, *Analytic Arithmetic of Algebraic Function Fields* (Dekker, 1979).
- [6] A. Knopfmacher and J. Knopfmacher, Arithmetical semigroups related to trees and polyhedra, to appear in *J. Combin. Theory Ser. A*.
- [7] G. Pólya and R.C. Read, *Combinatorial Enumeration of Groups, Graphs, and Chemical Compounds* (Springer, 1987).
- [8] D.H. Rouvray, Isomer enumeration methods, *Chem. Soc. Rev.* 3 (1974) 355–371.
- [9] Z. Slanina, *Contemporary Theory of Chemical Isomerism* (D. Reidel, 1986).
- [10] N. Trinajstić, *Chemical Graph Theory*, 2nd ed. (CRC Press, 1992).
- [11] I. Ugi, J. Dugundji, R. Kopp and D. Marquarding, *Perspectives in Theoretical Stereochemistry*, Springer Lecture Notes in Chem., Vol. 36 (1984).