Intellectual Properties Rights: Open Access
ISSN: 2375-4516

Robson, Intel Prop Rights 2013, 2:1
DOI: 10.4172/2375-4516.1000108

Review Article Open Access

# The Concept of Novel Compositions of Matter: A Theoretical Analysis

**Barry Robson***

*University Director of Research, St Matthews University School of Medicine, The Dirac Foundation, Oxford shire UK and Quantal Semantics Inc., North Carolina, USA*

**Abstract**

Here is discussed in the manner of a review the nature and uses of information measures in the discipline of patenting. From one perspective, the information content in a patent diminishes rapidly as the broadness of the claims increases. Claims made by Markush representations facilitate the quantification of that. The equations will approach yielding zero information if a massive number of chemical themes were implied. Importantly, a more detailed examination of these equations have implications that allow discussion of various aspects of novelty, reasonable consistency with a specific purpose, and perhaps even how many arguments and counterarguments there should be between examiner and assignee.

**Keywords:** Novel compositions; Patents; Similarity

## Introduction

### Information from patents

By serving to protect intellectual property in exchange for disclosing new insight and data for the benefit of science and engineering, patents are well recognized to be a rich information source [1]. Patents are also of value to science by helping define what is potentially novel by what is *not* covered, and conversely also preventing wastage of time and effort in "reinventing the wheel". To facilitate all that, computer-based *patent analytics* has developed to enable automatic extraction of useful information from patent text, including chemical formulae [2-6], their (typically protein) targets whenever they are potential pharmaceutical agents or of concern as toxic [7-9], as well other useful content. Provision of software and associated services relevant to patent analytics is an industry in itself [10-13]. the main problem is the relevant, prevalent, and perennial one of what is meant by the similarity of compounds. The general discipline tackling these and related issues is often called molecule mining [14] for a comprehensive bibliography, and [15].

### Novelty and scope

Fresh to the scene, one might think that "novel compositions of matter" as molecules and materials that have not, to our knowledge, previously existed in nature, would be a relatively straightforward concept. However, the force of the word "*novel*" we mean "*dissimilar* to that which has existed before", and by "*dissimilar*" we mean "differing in a non-trivial way". That is, differing in a way that is not obvious in the sense that, say, replacing butanyl by a pentanyl side chain seems obvious, or in some cases replacing an atom from another atom with similar properties, from the same column of the periodic table, say iodine by bromine. Such considerations remind us that the natural use of chemical formulae to address such matters is not the main mission of the pharmaceutical industry. Drug discovery is traditionally dominated by laboratory chemists, and so the molecule is indeed seen as a formula, a drug candidate being partly determined by area of expertise and ease of synthesis. It is, however, well known that even a change of a single atom in a large molecule to that of a different element can dramatically eliminate or change the biological action, whilst two molecules with completely different chemistry might, by having the same Vander Waal's and polar or electrostatic surface, perform exactly the same biological function albeit possibly with somewhat different effective concentrations. It is all because such a surface represents one half of the story about a continuous field of energy interaction with a biological target molecule that provides a complementary surface (albeit influenced by the surrounding environment); if favorable, it allows specific binding. In principle, the focus on formula hugely helps the discipline of patenting. It reflects the fact that adding, removing, changing, or rearranging atoms in a molecule occurs in discrete jumps, computationally a matter of integers employed in a clear cut graph-theoretic description (in contrast, the above field involves continuous values that are difficult to compute realistically even if we did know the target).

With a Markush representation [6], the combinatorial possibilities can be so huge that the problem can still feel like a continuous blur. "There are millions of combinations of side groups" is often said to be a common complaint from the examiner in the patent office in regard to a claim. One has the sense that patents for novel compositions of matter claim regions of prior art as fields of scope [16] in a huge chemical representation space of all possible molecules, and that between these regions lie the "white spaces" where the fields are so weak that they are free to be explored. Assignees generally try to increase the fields of scope whilst the patent office applies the art of claim limitation, i.e. applying intra-claim restrictions in an attempt to reduce them to a reasonable range with a crisply and reasonably defined boundary that leave no room for dispute in a court of law. Nonetheless, "reasonableness" is a concept that requires significant chemical experience, and "chemical experience" is by definition a troublesome concept when dealing with novel chemistries.

### Scope and utility

In a 2011 study [17], the author and colleagues applied somewhat unusual molecule mining techniques and drew the conclusion that "*chemical similarity and novelty are human concepts that largely have meaning by utility in specific contexts. For some purposes, mutual information involving chemical themes might be a better concept*".

***Corresponding author**: Barry Robson, University Director of Research, St Matthews University School of Medicine, The Dirac Foundation, Oxford shire UK and Quantal Semantics Inc., North Carolina, USA, Tel: 1-345-928-7242, E-mail: robsonb@aol.com, brobson@smu.ky

**Received** June 26, 2013; **Accepted** November 11, 2013; **Published** November 15, 2013

**Citation:** Robson B (2013) The Concept of Novel Compositions of Matter: A Theoretical Analysis. Intel Prop Rights 2: 108. doi:10.4172/2375-4516.1000108

This is explained and extended in this paper. However, it should be immediately stated that what was meant by "utility" in the above was not utility of the candidate drug molecules in treating or preventing disease. Rather, it was addressing what makes most sense for similarity assessment algorithms as regards their usefulness in selecting and ranking candidate compounds. This turns out to be insightful, nonetheless, in regard to readdressing the concepts of similarity and novelty. The initial aim of our project was to provide complementary tools to support patent based chemoinformatics systems developed by our colleagues [15,18]. The overall study with IBM colleagues involved using very high performance computing to read all US patents at that time, and to analyze a patent data base generated consisting of 6.7 million compounds re-expressed in SMILES codes [19] as character strings that represent the chemical formulae of compounds, alongside assignee and patent reference. In this case, the data base records were seen as quite rich, i.e. with many factors, because the SMILES formulae typically described, of course, many chemical themes in many combinations. "Analysis" referred to determining empirically what is meant by a theme, and the association constant between them, and with assignees (primarily chemical and pharmaceutical companies). Such studies give insight to what assignees and patent officer appear to agree to see as the reasonable scope of a patent in general, and the notions of similarity and of novelty.

### Experiences of utility of similarity testing algorithms

When undertaking the previous study [17], applying rigorous substructure similarity tests to very large collections of SMILES strings, and especially for similarity between many members of the collection rather than a single query, can be a rate limiting step for larger molecules in a workflow. We originally sought to speed this by first using a highly optimized industry standard for pattern matching, the regular expression ("regex") [20] in the context of the Perl language [21]. Obviously, if two substructures contain even one different specified element or different numbers of each type, they cannot be an exact match to a graph method as an exact algorithm, and this idea can be extended to finer details without invoking a full graph method. It was then noted that more exact graph-theoretic similarity tests were being pushed further and further down the workflow and still achieving overall utility. Many other workers have had similar experiences, or expressed similar ideas in expectation of such experiences, though not usually or necessarily concerning regular expressions specifically; [22-32] for an example bibliography of earlier work, and particularly a study on "*matching-relations, user-defined match levels, and transition from the reduced graph search to the refined search*" [32]. This idea of increasing levels of rigor of testing which catch more and more compounds as in some sense similar, when at first examination they are not so, is important in exploring the notion of similarity. However, in order to cover a set of molecules which the assignees sees as essentially similar, patents themselves contain the above-mentioned generalized descriptions of novel compositions of matter called Markush representations [6,22,23]. Searching them to see if a composition of matter is novel is a non-trivial task. Similarity is in large part a combinatorial problem, of what combinations and combinations of extensions to the description of a compound remain essentially similar. Similar combinatorial issues arise in many other chemical contexts [24-32]. This is examined from an information perspective.

## Theory and Methods

### General overview

"Methods" are included for brevity because the algorithms and hence software are essentially little more than basic information-theoretic formulae that quantify the following considerations; further required theoretical and methodological details are given in [17], but enough explanation will be made available to understand the principles. The considerations are, in the first few Sections, some different but interrelated perspectives of how much information a claim for a novel composition of matter should contain, and what that means. Formulae are presented in later Sections. Of particular interest are Markush representations in chemical patents, because there information takes on a tangible form which is much harder to quantify in other kinds of patent, even though the same ideas are often applicable in principle. Like any broad description covering many possible specific embodiments, a Markush representation is not specific. It formally contains less information than just one exact chemical formula as a sole embodiment. A fuller discussion on this requires the notion of substructures interpreted here as "chemical themes". In addition, the formulae developed then suggest broader philosophical issues of patenting that are not necessarily intuitive.

### Plethora of heterogeneous material

Combinatorial Explosion in the Markush Representation: A Markush representation is a formula stating that an $i^{th}$ group $R_i$ "can be any one of…" and there may be several such R. In what might be considered a moderate case, we might have something like a mere 5 different possible replacements at each of just 5 possible sites. That is $5^5$ = 3125 different molecules. As far as the discussion in the patent goes, a Markush description reduces from a $5^5$ = 3125 to $5^2$ = 25 problem, but it remains 3125 chemistries. It is unlikely that all would be reduced to practice by synthesis and testing in traditional ways, unless the proposals were generated experimentally by some kind of combinatorial chemical method. The example of 5 replacements must be, scientifically at least, an understatement because each chemical group may imply a new set of sites for replacement. If is again 5 then we have $(5^5)^5 \approx 10^{17}$. Claims limitation is discussed in the next Section, but it is clearly needed. Whatever the working rules of a nation in regard to patenting, it seems hard to accept a scope of patent that would seem extremely implausible to demonstrate by reduction to practice by synthesis for all molecules intended to be covered, and arguably unreasonable where there might be millions of distinct applications for the variety of molecules intended to be covered, say as distinct pharmaceuticals with distinct targets. This is distinct from new methods of synthesis, which could affect many future syntheses, but improve on and go no further than the process of syntheses itself. Armed with modern combinatorial methods, $10^6$ is probably more like a realistic number of molecules for a pharmaceutical company to manage per year, currently. In theory one can imagine that the above example of $(5^5)^5 \approx 10^{17}$ might be reduced to practice by some nanotechnology-style approach. It would represent a mass of about 0.1 mg of different molecules of molecular weight around 600, and if requiring synthesis and detection machinery of about the same size as an enzymological complex (such as the Pyruvate Dehyrogenase complex), then one might accept the idea of a flow-through screening path capable of detecting specific complexes to deliver the data, but it would still require perfect sensitivity to detect individual molecules. Allowing duplication of molecules to overcome that, a future nanotechnology-style industry might be able to handle a Kg. However, taking $(5^5)^5$ "one step" up to $(6^6)^6 \approx 10^{28}$, the representational mass becomes 10,000 Kg, and for $(7^7)^7 \approx 10^{41}$ becomes $10^{13}$ Kg of distinct molecules, roughly the mass of Mt. Everest.

### Claims limitations for markush representations

By re-expressing the implications of such numbers in the analogy of a Markush representation in simpler cases, unreasonable scope can be revealed without appeal to the above "unfeasibility of synthesis" notion. The $10^{41}$ seems unreasonable because $10^{41} \approx 20^{32}$ would be analogous to a Markush-style claim "where X is a polypeptide chain of 32 amino acids where each amino acid is one of the naturally occurring 20 amino acids found in proteins". 32 is also an interesting number. It is roughly the size of a smallest protein or smallest domain of a protein (that was probably ancestrally a protein in its own right). If we had powerful enough tools for protein design, that might constitute many millions of proteins with different, say pharmaceutical or enzymic (catalytic) applications. That seems unreasonable in a patent claim, and it would seem to suggest that the assignee was more interested in "claiming territory" in the patent chemical space than reducing claims to practice and exploit them. It is presumably this general kind of reasoning, and not least the difficulty of searching many patents that use very broad Markush specifications, that suggested confining the power law to reasonably small numbers of alternative substitutions on a single core or framework recommended in the Federal Register on August 10, 2007. Notably, an intra-claim restriction is proper if all species share a feature that is substantial and essential for common unity, and Markush alternatives must be substitutable, may not encompass other alternatives, may not be a set of further alternatives, and must not make the claim difficult to construe (it leaves open the question of what "substantial", and "difficult" mean). Strictly speaking then, the $10^{41}$ is seen as more digestible if there a shared substantial and essential feature, usually seen as a core. In the following is considered that more generous perspective, and indeed a little more generous, because it seems unreasonable to consider an "integral feature" where variations of it are of the same nature as by which we consider variations within the R groups. It may be said that this allows a fair claim of essential chemical theme. It is not necessarily a fair claim by the plethora of heterogeneous material (see above) or, related to that, by the plethora of applications (see above) that it might imply. A "fair claim" evidently does not necessarily mean that an unfair claim is an unacceptable one, because as discussed in above, the number of $10^{41}$ implied compounds can pale in comparison to the scope of some published patents.

## Chemical themes

Behind the patenting scenes there is the notion that the context of the heart of the invention, typically seen as a core, does matter, else one could simply specify a molecular structure such as a core without the groups R that the Markush representation introduces. Rather than that extreme, however, there is the notion of reusable component parts, in which structures can be reused, but not entirely freely because there are some concerns. A frequently quoted example of innovation or innovation is the first repurposing, an event lost to history, of the flower pot as a chimney pot [32]. As pottery or ceramic the flower pot was effectively a novel composition of matter, but its radically different use did not represent a new novel composition of matter even though it was bound strongly to the chimney by mortar. In many respects a piece of molecule that can be conceived as a more or less distinct object and reappear in different molecules raises the same issues as the plant pot. However, the picture can be blurred. The situation in chemistry can be more akin to the clay or mortar components diffusing into the pot and modifying its nature. Electrons are matter, but more manifestly demonstrate quantum mechanical laws than do nuclei. That is, at least in regard to being fickle as to which atoms, and hence sub-fragments, that they belong to. It may be considered a context dependent effect, meaning that the electronic effects of the surrounding structure can have significant chemical effect on a substructure capable of existence as a molecule in its own right. That is, allowing for addition and removal of hydrogen atoms: hydrogen atoms are not explicit in the study, they are there by implication to satisfy free valencies [17]. The problem is to some extent a point of view, but not totally. A biochemist would consider adenosyl diphosphate and adenosyl triphosphate as very different, but then so would a quantum chemist. Nonetheless, the simple fact that it is possible to identify substructure or fragments as recurrent chemical themes and associate them differentially with companies [17] indicates that the notion of distinct chemical themes as modular components makes sense. Our previous study involved automatic generation of taxonomy of chemical themes, because a question of what is a theme is formally a matter of whether we pool distant, close, or immediate family members. A chemical theme per se corresponds, by analogy with biological taxonomy, to a *genus*, and specific compounds correspond to species. A genus is a first order of generalization, and essentially Markush-like, but no attempt was made to intervene and enforce an arbitrary fixed core. The details are essentially as described in Ref. [17], but the main features may be stated briefly as follows. Apart from C, N, O, S, and P keeping their identity, all atoms in the same row of the periodic table were as very first step pooled into one atom type selected as of most interest to chemical (not necessarily pharmaceutical) companies. Hence one would either see C which really is carbon, or Si which is not necessarily silicon but any other member of the same column of the periodic table. In consequence there are 12 typical bonding atom types of rows 5-9 of the periodic table which tend to occur in practice. X* such as e.g. C* is used to indicate a run two or more of the same atom (element) bonded together. Double bonds '=' and triple bonds ' #' are retained in representation although in side chains connecting to the core by double or triple bonds, the side chain is only retained as e.g. (=C*) or (#C*). Rings structures with chemical themes are preserved, and the numbering system showing the number labels associated with atoms to show the join, e.g. atom labeled 6 to atom labeled 6, are reset to start from 1. If the ring is *"snapped"* by extraction of the chemical theme, however, the surviving number indices are retained as information that a ring was present.

How many distinct chemical themes described in such ways make up the patented chemical space? The detailed description of chemical theme generation in Ref. [1] which starts with chunks of 35 SMILES characters lead to fragments of 9-13 "pooled atom" symbols such as C* of 41 types. In theory, then, one might estimate about $41^{13} + 41^{12} + \ldots 41^{9} \approx 10^{23}$. Most however are impossible like O#O (triple bond) or F=F within a larger molecule, or implausible like N-N-N-… with $20^{13} \approx 10^{17}$ as a more reasonable upper estimate, and the comprehensive CAS registry (see Results) holds less than $10^{9}$ organic and inorganic compounds of any size. The author estimates that the number expressed for the above chemical themes occurring in the registry would be more like $10^{5}$-$10^{6}$. Since one may avoid uncertainties by addressing how much a new claim would if accepted add to the patented chemical space of industrially potentially interesting molecules, then note that some $6.7 \times 10^{6}$ distinct compounds can be extracted from mention in all US patents (of circa 2008) which, by any reasonable criterion of chemical theme in the above spirit, contain about $10^{4}$-$10^{5}$ such. As discussed in Results (but needed now and used below), the author estimates that there are in the patented chemical space close to $10^{5}$ distinct chemical themes from a chemist's perceptive, consistent with the earlier work [17]. This estimate is conservative in that only US patents are considered, but is generous because it takes account of immediate substructure context as discussed later below, not just what is in the chemical theme itself. It should therefore be stated up front that the practical use of such a collection for considering specific patenting

cases is limited because that number represents an average of likely many overestimates and underestimates of how affected a substructure might be by its surrounding context. As an average, and a recurrent result using different and rather arbitrary criteria, the number $10^5$ is seen as an important "constant" in the following considerations, although one notes that it is the principles remain important if this number is reappraised.

## Information content of a markush-based claim

There is a simple means to quantify the scope of its contribution to the known chemical space, at least as rough estimates of the numbers of bits (meaning here specifically the binary units) of information involved. Note recall that $\log_2(n) = \log_e(n) \times 1.442695... = \log_{10}(n) \times 3.321928...$ A claim by a Markush representation A that encompasses n[A] specific species which is less than one B that encompasses n[B] carries more information. The extra information it carries is $\log_2(n[B]/n[A])$ bits. It is useful to get some idea of an absolute value of a reasonable claim as opposed to such relative values. The previous study [17] identified some $10^5$ chemical themes of recurrent interest to the chemical industry of very roughly comparable molecular surface area, and these themes themselves contain a code that is essentially a roughly analogous to a Markush representation. Indeed, a typical Markush description can represent it, and in any event some kind of Markush description can. A patent that would, if awarded, add a reasonable and comparable Markush description of a new chemical theme to the patent chemical space thus represents 1 in $10^5$ of the chemical themes discussed above, or about 17 bits, the information required to pick it out of the chemical space first time compared to a random selection. A reasonable and modest claim obeying the above claim limitations might be 1 core × 100 R1 × 100 R2 × 100 R3 = $10^6$ distinct compounds which if accepted into the patented chemical space, would raise its information content from $\log_2(10^5) \approx 17$ bits to $\log_2(10^6+10^5) \approx 20$ bits, but only if the $10^6$ distinct compounds represented $10^6$ distinct chemical themes, and a major point of the above claims limitation principle is that there is in the Markush representation only one new chemical theme by the more usual concept of a chemical theme – the core. A claim with just one such core adds one new theme to the patented chemical space, hardly changing the number of themes in it, so it is thus worth 17 bits and not, it is important to note, 20 bits implied by the $10^6$ distinct compounds defined by the Markush representation.

## Excessive and modest information content of markush-based claims

A claim of excessive scope including what might be considered as more than one distinct chemical theme will reduce this value. 32 for the number of amino acid residues mentioned earlier, in that context seen as an unreasonable claim, implies $20^{32}$ yielding 136 bits. To have some idea at least an upper limit to the kind of scope that pass examiners but might be considered by some as ambitious, note for example: "*A Markush claims cover a wide series of possible compounds. Sometimes, the series may encompass billions or trillions of variants; sometimes the series is unlimited. For patent EP 0 535 152 … , for instance (by no means the more complex Markush patent), I have calculated (perhaps conservatively) that the minimum number of compounds covered by the literal wording of the patent could be 10 followed by 60 zeroes*" [33]. This means $10^{59}$ distinct compounds and $\log_2(10^{59})$ is approximately 196 bits! However, if that meant $10^{59}$ chemical themes added to a chemical space of a mere $10^5$ would mean that the patent carries some $\log_2((10^{59}+10^5/10^{59}) \approx 0$ bits of information required to pick all those chemical themes out of the updated patented chemical space. In effect, the patent would contain almost no information. About 17 bits seems a reasonable an upper limit for a Markush-based claim. Of course, if $10^5$ were a hugely gross underestimate for the patented chemical space by many orders of magnitude, and then corrected, a very broad claim would seem less outrageous. If claim were allowed 10 distinct cores and hence 10 Markush representations, $\log_2(10^5/10) \approx 13$ bits. What constitutes a core is an arbitrary human perception, and of course we can imagine it as containing a Markush group R but something more like allowing an extra carbon atom within a core ring of carbon atoms, or an extra oxygen atom next to one already present in the ring. This is much closer to the notion of a chemical theme used in Ref. [17] and the current report, as described above. By inspection of sampled patents, 13 bits implied by a claim is a typical modest claim but such modesty does not seem common amongst claims. $2^{13}$ is just 8192 distinct compounds, and a very modest coverage by the standard of the kind of example discussed above. Recall that a broader claim means less information: increasing the number of themes covered will drop it below 13 bits. Increasing that information by being even more specific is a noble idea, but the range 13-17 bits does not give much room for maneuver. We might ask whether 13 bits seems reasonable for a Markush based claim on other grounds, as follows.

## The comparable biological information in a molecule

Recall the ambitious claim "where X is a polypeptide chain of 32 amino acids where each amino acid is one of the naturally occurring 20 amino acids found in proteins". It would seem proper if a Markush-based claim for a bioactive molecule contained roughly the same amount of information (in the above defined sense) as a molecule needs to exert some kind of biological action. In both cases one is looking at the information implied in one compared to all the possible alternatives available. Numbers of bits in the vicinity of 13-17 are interesting from the point of view of information required for a drug molecule to recognize its protein target and *vice versa,* effectively the information that is the real focus of the pharmaceutical industry. Such a range is the ballpark in which both drug discovery and patenting are played. Recall again the polypeptide of 32 amino acid residues. Only a part of it is involved in the molecular recognition that dictates the information required if we are to consider a specific potential drug for a specific target. We are interested in weaker recognition that could still constitute a biological action. Consider that a binding site as a recognition site might contain just 5 types of strongly distinct group, say non-polar, hydrogen bond donor, hydrogen bond acceptor, positively charged, negatively charged, and there were 5 to 8 such. The range $5^5$ to $5^8$ gives approximately 11 to19 bits, and so with 13 as not unreasonable. Such estimates can be subject to various criticisms and so it is fortunate that protein binding sites on DNA give a more direct indication of information content in protein-ligand recognition. They are typically 4 to 30 base pairs long (8-60 bits). However, a typical strong binding sequence such as 5'GCGTGGGAGT3' [34] is about 10 base pairs long (20 bits), and can typically undergo substantial mutations without losing binding activity, with some 2 strong determinate bases and many contributing weaker ones [34], suggesting that approximately 13 bits is reasonable. That it is more than just analogy is suggested by the following.

## Molecular versus digital libraries

The relevance is that mathematically there is not much difference between a digital library and a molecular library despite the very different nature of the hardware, and so of the fact that the protein target is the query in the molecular library case. Another approach which perhaps highlights this is to say that our set of 32 amino acid proteins would also have bioactivity as many different immunologically

active and distinct variants. It is well know that an epitope is a region of roughly 5-15 amino acid residues. Approximately 5 amino acids represent the molecular recognition patch of Vander Waals and dipole or electrostatic forces fairly typical of a small binding site (in the sense of on the protein or the ligand). It suggests a minimal requirement of $20^5$ distinct pentapeptides, recalling that many of these are also immunologically distinct epitopes, i.e. antigenic determinants multiple "functions" in regard to molecular recognition. However, the amino acids are not equally similar, so that one may replace one by about two others ("conservative substitution"). That gives approximately $7^5 \approx 10^4$, about 13 bits of information. However, as many as 9 residues in a single binding site of a protein can have atoms within 15 Angstroms of atoms in the complementary binding site of, say, an antibody so there is an upper limit per site of around 23-25 bits for a strong recognition interaction. The antibody, MHC protein etc. protein as target now represents the query. At first site an apparent discrepancy is that association constants between proteins and ligands are typically in the range $10^6$-$10^{12}$, i.e. implying 20-40 bits of binding information with the higher values typical of epitope-antibody interactions, and candidate drugs of pharmaceutical interest associations constants in the range $10^8$-$10^{10}$ i.e. 26-33 bits. However, this is relative to associated and dissociated form without competition, a library in which the epitope and the solvent are the only competitors to satisfy the query. The library of interest is arguably a conceptual mix ideally including all the other ligands that have specifically described as prior art, which is unfeasible and probably with surprising consequences if it were feasible. There is a case that it should be a mix of molecules with the claimed Markush core but with R groups replaced by specific chemistries that are not covered by the claim, but this has several difficulties. Fortunately we can obtain a sense of the discriminatory power by subtracting the largest corresponding information for competitors in such a mix. With 20 bits for the weaker binding in the range 20-40 bits and the 33 bits for the upper value for drugs of typical interest, 13 bits emerges as a reasonable value.

## Quantifications of surprise

We now try to give more formal rigor to assessing and applying information. One position that we can take on patenting is that if there no surprise in the claim given prior art, it is not novel. Above, we took a specific interpretation of this, i.e. how surprised we would be to pick it at random from the patented chemical space if it were added to it, and hence how much information would be required to pick it first time. Properly speaking, however, it is necessary to consider extension to the case of sparse data because a truly new invention appears just once. In practice it tends to be "*of its time*" and precipitates related compositions of matter as spikes in time of closely associated chemical themes, but nonetheless it is sparse data, at least in the early days of research, development, and exploitation. The theory of information from sparse data was formulated by the author in the field of bioinformatics [35,36] and later for clinical and pharmaceutical applications [37-40]. Mutual information is just one special case of a surprise measure, part of a broader *zeta theory* of data mining, management, and inference [41]. More generally we can write

$$\text{Surp (A; B; C ;...)} = z(s, o[A, B, C...]) - z(s, e[A, B, C...]) \quad (1)$$

Here o[ ] and e[ ] are the observed and expected frequencies, i.e. number of occurrences, as discussed in the next Section 2.10. Above, the result of integrations of metrics over Bayes posterior distributions [36] has yielded measures of surprise in terms a linear expression in functions that are summation series. They can be generally expressed as the Riemann zeta function ζ(s) as z(s, n), incompletely summated up to n rather than infinity. Function z(s, n) is definable for real and complex values of s and n but is simple for natural numbers k = 1, 2,3,...n; results can be found by smooth interpolation of the following: $z(s, n) = \Sigma_{k=1,2,3,...n} k^{-s} = 1 + 2^{-s} + 3^{-s} + ...n^{-s}$. For s=1, Equation. 1 is expected Fano mutual information I (A; B) using the natural logarithm ln( ) [36]. More precisely, there is a correction by a small Euler-Mascheroni constant γ = 0.5772… to consider, but not if we take the zeta functions as axiomatic of the information available to the researcher through the data, and in any event it effectively cancels in subtractions between zeta functions [42], which is the usual use case like in Equation.1. For that reason we use here the natural and not binary logarithm here, so that the information units are nats not bits (recall $\log_2(n) = \log_e(n) \times 1.442695...$). Traditionally and equivalently for large data n→∞, I (A; B) is the log of the association constant K(A; B) = P(A & B) / P(A) P(B), and relates to free energy kTlnK(A; B). The concept of an alert is similar. It is specifically reserved as meaning a change in surprise, i.e. of the measure Surp, typically with time, and typically for different time periods, says changes in the current year as opposed to previous years: $\Delta\text{Surp(A;B;C;...)} = \text{Surp(A;B;C;...t}_2\text{)} - \text{Surp(A;B;C;...t}_1\text{)}$ [17]. Examples of other choices of s include s=0 (at least for one branch of interpretation below s=1), whence the above summation as written means that z(s, n) = n. When s → ∞, Equation. 1 very rapidly approximates (i.e. for quite small values of n as o[ ] or e[ ]) +1, 0, or −1, tertiary logic.

## Observed and expected frequencies and patent information content

By data we mean that in Equation. 1 arguments o[ ] and e[ ] as observed and expected frequencies (counts) of states, events, measurements or descriptions A, B, C... Note that the function is also valid for z(s, n=0), as when o[ ] = 0 or e[ ] = 0. No data implies no information: these functions measure information about a system, but pragmatically only that accessible to the user via the data. Expected frequencies can be interpreted as in the usual statistical sense and as used in the chi-square test, such that $e[A,B,C,\ldots] = N^{1-n}\,o[A]o[B]o[C]\ldots$ for n items A,B,C etc… given total amount of data N. However, Equation.1 applied to the calculations of the information in a Markush-based patent in preceding Sections would be interpreting e[ ] as o[claim], the number of claimed new chemical themes and the original o[ ] of Equation.1 as the chemical patent space with these claims admitted to it. There is another way and equivalent way of looking at the required measure. As discussed below, it is valid to add virtual frequencies as prior belief, but also this includes including actual frequencies from prior studies. We think now of o[patented chemical space], currently $10^5$, as the patent space prior to accepting it.

$$I_{claim} = z(s=1, o[claim] + o[patented\ chemical\ space]) - z(s=1, o[claim]) \quad (2)$$

In the current preferred ("examiner-friendly") claim, o[claim] = o[Markush cores], and ideally o[Markush cores] = 1. Recalling the Euler-Mascheroni constant from above, we can think of the result that the "ideal claim" has $I_{claim} = 1.442695 \times (\ln 10^5) + g - 1)$ bits which comes out as much closer to 16 bits instead of closer to 17 bits, using the "finite data" model that the zeta function implies. Conversely, if the patent implied a huge number of distinct chemical themes, o[claim] >> o [patented chemical space], then $I_{claim} \approx 0$.

## Weight of evidence

Assessing patent applications is evidently a job that takes significant skill and expertise, suggesting that an Expert System of some kind would be appropriate. When also clearly a decision support system, a

decision theoretic approach is appropriate, and the above theoretical considerations remain relevant. The filing of a patent implies the assertion, say A, by the assignee that there is novelty, perhaps given certain arguments B, C... as conditions that support that. The patent office does its due diligence to try and disprove that assertion, by providing evidence for the negative or complementary case ~A. In principle, one match with prior art would be sufficient to clinch the matter, but as shown above, the issue can be more complex. One way to look at this is as a decision process, based on log predictive odds. A "bonus feature" may be introduced a this stage. Sometimes it is not possible to obtain a weight of evidence as a count, i.e. a frequency of observations, or there is an aspect of that which we wish to include along with a count, as a prior belief. However, we may at least still be able to quantify some kind of expected frequency e[ ]. The theory of probability distributions that gave rise to the use of expected information as zeta functions [36] demonstrates that the consequence is simply as if we add virtual frequencies, here e[ ], to the natural frequencies o[ ]. It comes from the prior probability distribution, say as a binomial distribution $P(A)^{e[A]}(1\text{-}P)^{e[\sim A]}$. Overall, the assignee provides information z(s,o[A,B,C,..] + e[A,B,C]) and the patent office seeks to provide the counterevidence z(s, o[~A,B,C]+e[~A,B,C,..]).

$$I(A: \sim A \mid B; C;..) = z(s, o[A, B, C...] + e[A, B, C]) - z(s, o[\sim A, B, C...] + e[\sim A, B, C...]) \quad (3)$$

The weight of evidence is said to be in favor of novelty if the above is positive, and against if negative. We should really include all four pillars of evidence that are brought together in an odds ratio, the log of which can be shown to be I(A: ~A | B; C;...) - I(A: ~A | ~(B; C;...)). However with a number of conditional arguments B, C, D and their collective negative or complement, the log odds ratio is is closely approximated by Equation.3, because P[~(B; C;...) ≈ 1.

## Unpredictability as novelty

A simple patent application of measures of this kind is that if A, B, C, etc. are chemical themes and a claim is made that a structure (A, B, C, D) is a novel combination, it is not so with respect to (A,B) with (C,D) if I(A, B; C, D) ≈ 0 in the patented chemical space, and it is not so if with respect to (A,B,C) with (D) if I(A, B,C; D) ≈ 0, and so on. Also, for similar reasons, if I(A; B, C, D) ≈ I(A; B) + I(A; C, D), then (B) and (C, D) are purely randomly associated in the context of A. A simple way to encapsulate all these possibilities is to say that if we cannot make a reasonable prediction of something c from the patented chemical space, it is novel or not obvious. This Section describes the basic ideas for future decision support systems in patenting; however, the theoretical outline required seems fairly clear. For a simple decision process or prediction of case c based on many A, B, C, etc. we may think in the following manner. The problem is that anything other than the most trivial cases of a single I(c: ~c; A) as a metric (log predictive odds) will be an argumentation model, i.e. involving many contributing terms as weights of evidence. They provide estimates of I(c: ~c; A, B , C , D , …) as the joint information using simpler terms such as I(c: ~c; A), and I(c: ~c; B | A) (representing "given A", i.e. removing the contrition of A to avoid adding it in twice). However, any multifactor term could be expanded in different ways, any of which would alone in principle suffice:

$$I(c: \sim c; A, B , C , D , \ldots) = I(c; A , B , C , D , \ldots) - I(\sim c; A , B , C , D , \ldots)$$

$$= I(c: \sim c; A) + I(c: \sim c; B \mid A) + I(c: \sim c; C \mid A , B) + I(c: \sim c; D \mid A , B , C)\ldots$$

$$= I(c: \sim c; A) + I(c: \sim c; C \mid A) + I(c: \sim c; B \mid A , C) + I(c: \sim c; D \mid A , B , C)..$$

$$\vdots$$

$$\vdots$$

$$= I(c: \sim c; B) + I(c: \sim c; A \mid B) + I(c: \sim c; C \mid A , B) + I(c: \sim c; D \mid A , B , C)....$$

$$= I(c: \sim c; B) + I(c: \sim c; C \mid B) + I(c: \sim c; A \mid B , C) + I(c: \sim c; D \mid A , B , C)...$$

$$\vdots$$

$$\vdots \quad (4)$$

Recall that estimated as zeta functions from actual data, we may not get exactly the same result using different terms and neglecting different terms. We would like to include all that we can get. The actual way to do this and use it is fairly straightforward in the case of patients, if not patents. Suppose that we examine a patient record as a query, and compare such information measures derived statistically from many patient records as quantitative "rules" in order to predict a diagnosis, best therapy, prognosis or risks for a patient. If any rule as an expansion term matches the record of the patient under consideration (match of all of the r factors on the record), we add the information in favor of *c*, and if a term in any way does not (at least one of the r factors does not match), we subtract it. If there are any "don't knows", then no information is added or subtracted. We should not apply the terms of Equation.3 as rules with equal weight. It is easy to see that if we only allow the simplest terms I(c ; A), I(c ; B) etc then for n factors A, B, C,…used in prediction we have for each the weight 1/n. The weighting for terms as rules in general is arguably the combinatorial expression b(n, r) = + (n−r)! r! / n!, but not necessarily. If we actually only have N examples of rules with n factors, it makes no less sense if the weighting factor is simply 1/N.

In practice, as for the medical example, this requires a decision support system of some non-triviality. Indeed, research needs to be done, not simply development. Some possibilities are as follows. Of several ways to use this in the patent context, the simplest is that c: ~c is the claim. In that case, it is sufficient to think of I(c; A, B, C, D…) and remove ': ~c' throughout. More correctly, though, we should indeed think of ~c too, of c as "novel", and ~c as "not novel" for say A, i.e. already existing in the patented chemical space. There is a conceptual difficulty if, as earlier above, we think of the claim as if it were added to the chemical patent space in order to determine information measures, even though we must distinguish it a *c*, the claim under consideration. With one claim in a patent (there can of course in practice be more, even if it conceptually undesirable), and (it is hoped) just one prior claim A preexisting in the patented chemical space, the use of the zeta functions implies 1 – 1 = 0, but in this case we do not mean that the claim could not be predicted, but that its novelty could not be, i.e. the weight of evidence for or against it being novel is zero. In effect, the claim must be discarded on the grounds of "case not proven". However, with or without that particular approach, the usual case is that we cannot be certain that A alone suggests lack of novelty, but considering many other aspects may improve that. The simple and typical case is that A as Markush representation overlaps with the claim, but does not contain all of it, and so may B, C, D, etc.

## The claim as an assertion to disprove

Estimates of chemistry patents allowed have varied from

around 45%-76% in various countries, so a priori the probability of acceptability are around 0.45-0.76 and hence arguably roughly 1 bit. There is a case, however, for saying that at the moment of filing, even that is excessive and in fact there is no information at all. Many inference systems like that discussed are of course framed in terms of conditional probabilities, but our counterparts of them in terms of zeta functions lead to a philosophical position of importance that may at first seem counterintuitive. It should not be ignored, because other interpretations support it. Our counterparts have to be of form

$$P(A \mid B, C,...) = e^{z(s=1,\, o[A,B,C,..]) - z(s=1,\, o[B,C,..])} \quad (5)$$

It is easy to show that as o[ ] increase, we rapidly approximate P(A | B, C,...) = o[A, B, C,...] / o[B, C,...], and as o[A,B,C] increases, the probability approaches 1. These are very satisfactory as the classical or frequent interpretation of a probability. However, whilst it was pleasing to find that less and less data means less and less information, zero information in Equation.5, when o[A, B, C,...] = o[B, C,...], means a probability of 1. It is well known that information is defined as I(A) = –lnP(A) and with 0 = ln(1) and so this is consistent, but it is subject to interpretation. As has been used already earlier above, probability of finding a state or event A by chance determines the amount of information that is needed to identify it immediately. Equation.5 is itself a specific interpretation: if you have no data, the best default position for using the probability P in inference is the choice P = 1. The primary and rather obvious example of what it means, and a case in which the nature of the conditioning events B, C, are not important, is that there should be a large number of attempts o[B, C,..] to find prior art. What is less obvious is that Equation.5 will not only start at zero but start to closely approach a constant value when o[B, C,..] exceeds 20. This is when z(s=1, o[A,B,C,..]) – z(s=1, o[B,C,..]) starts to approach ln(o[A,B,C,..]/o[A,B,C,..]) within about 5%, the exact amount depending on the numbers. The problem otherwise is that there is no halting instruction except finding prior art that clearly refutes the claim of novelty. The nature of the Markush representation, the complexity of patent wording, and the difficult notion of what makes a molecule similar to something else could make a comprehensive search very long. The easiest interpretation of what constitutes an observation that advances o[B, C] by 1 is the examination of a patent, or better still a chemical theme or comparable construct, that when using data mining do not obviously refute the claim without negotiation with the assignee. In practice, an examiner sends an applicant a non-final rejection, the assignee amends claims or presents counter arguments to the examiner, and the examiner accepts amendments or counterarguments. The interest here is in the arguments and counterarguments. The above suggests that there should ideally be at least about 20 arguments (and consequently 20 counterarguments) based on chemical themes that appear to be similar to each chemical theme claimed. There should really be only one invention claimed per patent but it could be argued that variations of a Markush core are not fundamentally different inventions. Should after 20 passes the examiner then feel that the claim cannot be refuted; there is reasonable evidence for novelty. Of course, if the examinee can only find, say, 5 arguments, and cannot accept at least one counterargument, the process halts, and the claim is rejected. The above suggests that the examinee should have about 20 distinct best shots in which he can accept the counterargument in every case. This is onerous, but some patents can take up to three years to accept or reject.

Equation.5 has many other applications, including evaluating the terms in Equation.4 for predictive inference. Other interpretations of probability that support the above notion of probability include the following.

## Popper's principle of refutation

It is plausible that the filing of the patent may be seen as assertion of novelty without great weight of supporting evidence, since the assignee might for various reasons be biased in failing to find prior art, and indeed the would-be inventor might genuinely believe that there is none, since the effort to do the research would be undertaken on the belief that there was no prior art. Karl Popper's principle of refutation of evidence [43], can be can framed here by saying that P(A) =1 is usually at best weak support for A existing or being the case, and it is usually only contrary evidence reducing P(~A) reducing its value, that matters. It relates closely the "Black Swan" issue, the point being that while we may assert that all swans are white on the basis of common experience, only diligent search for exceptions and failing to find them will prove it to be true, and there is no practical halting instruction to say when due diligence has been done.

## Risk

The Popper position becomes of practical importance in matters of risk: a *"black swan event"* means such as a terrorist attack on, or an earthquake in the vicinity of, a nuclear reactor. It pays to design on the assumption that such will happen, and certainly not to hold to the optimistic proposition that such events cannot happen. As a general statement, it is weaker than the Popper position because in general the risk may not be of an outcome that is significantly costly. Nonetheless, if either potential risk or cost is unknown, or their value cannot be quantified, such ignorance obliges us to play safe and assume the worst. The patent office is acting in the interest of the assignee, because if the patent does not represent an invention, there is significant risk to the further investment made by the assignee organization.

## Neglect of probability terms implies using them with probability one

Viewed as participating in an inference process, the notion of P=1 expressing ignorance seems particularly persuasive. In performing inference by purely multiplication of conditional probability terms, say as in a Bayes Net [44], the vast number of probability terms that are always omitted as seemingly irrelevant or of which we are ignorant would have the same consequences if we included them with probability 1. It is relevant for us here because it is the same situation discussed for Equation.3, when neglecting a term is the same as including it with zero information. There, as in a Bayes Net, it may well mean that it had too many A, B, C, etc for there to be no data to evaluate it other than as 0 which the zeta function of an amount of data gives for no data. It is really the only justification for such neglect. The more we drill into all the all the dimensions of "Big Data", the more we are likely to discover that this necessary neglect was unfortunate, but assuming 1 in absence of such an extended quest seems reasonable. Certainly, assuming that any one of perhaps trillions of terms ignored has a very low probability will dominate the joint probability of the network as close to zero. Picking an arbitrary number like 0.5 for every term would similarly give a very small and perhaps unreasonably small joint probability. Assuming P=1 is the safest bet for ignorance here, and Equation.5 demands that we agree.

## The hedge

When the scope of a patent is broadened, the assignee is hedging the bet that it covers the cases that will actually work, i.e. be useful. This seems different from information content as novelty, because it relates to scientific or industrial worth as the truth of the assertion. It is more likely that a claim hedged in the sense of being a broader claim is more

likely to be technically true as an assertion, meaning here assertion of the explicit or tacit usefulness of what is covered is likely to be the case. Nonetheless, it only appears to be a different interpretation when we think of P=1 as signifying a strong statement. The notion of P=1 expressing ignorance is consistent with the trend of the "hedge" notion that we increase the probability when we hedge or weaken a statement with a qualification like "sometimes" (because we have more confidence in it). For example, the categorical interpretation of a P(A|B) based on counting as P("all B are A") does depend on the assumption that there will be no errors or exotic cases in perhaps millions of observations that will throw "all B are A" to untrue. A hedge typically means that we can have a required smooth function of P(A) on n(A) but raise P to the power of the reciprocal of a positive integer, say j, which makes the probability larger. We can use the implication of a hedge like "for all practical purposes" by taking the square root of the probability (i.e. making it a larger value), a popular choice in Expert Systems. For the strongest possible hedge for the weakest possible interpretation of a statement we obtain P=1.

## Results

### Statistics of chemical themes

The patent data base of 6.7 million compounds when searched produced 833,333 genera detected by partial matching with regular expressions, reducing to 99,470 chemical themes of some 9-13 atoms as judged by the graphic methods (using connectivity matrices), i.e.17 bits. 13% of connectivity matrices represented 1-3 of the original genera, 1-2% contained 4-6, and 75% contained 7-28. These themes can overlap, and there is a great redundancy of sub-themes. Removing overlaps and identifying modular parts gives 30,132 distinct themes, i.e. 15 bits, mostly of about 9-13 non-hydrogen atoms. These author's later calculations with minor refinements are essentially consistent with the previous report [17]. However, based on more recent work there is good reason to reappraise this number and restore it close to the above 99,470 substructures distinguished on a graph-theoretic basis, as follows.

### Context dependency

This is the context issue discussed in Theory regarding how surrounding chemistry of a containing molecule can affect the properties by, for example, a highly polarizing atom or group in one context and not in the other, or a double or triple bond in one compound context and not in the other. Moreover, if we insist on the limit of breaking a highly integral core such as a steroid multiple ring cores into its component substructures, we should at least retain the original in our database of chemical themes. In the previous report [17] there was interest in how many failures to match by a regular expression (but which can be matched by a graph method) were actually appropriate misses in the above sense. Some human chemical expertise is required to estimate what fraction are in practice appropriate misses because of the neighboring chemistry effects. Simply searching intersecting rings and strongly polarizing groups and double or triple bonds connecting the substructure to the overall structure can be automated, and making allowance for the fact that polarization and bond order are matters of degree, then this increased chemical themes as now distinguished by their context some 2.6 – 3.3 times, making $10^5$ a reasonable estimate of the number of true chemical themes.

### Distribution of association of chemical themes with companies

This is mutual information of the form I (chemical theme; list of companies associated with it). At the top of the list 34 companies filed patents containing themes CN(C*1)C(C*(N=2)C(C)C)= NC2N3CNC* at 10 bits of association. That is some 935 times more than expected on a chance basis considering the occurrences of chemical themes and each company separately. The distribution approximately follows Zipf's law (See Results) but with occasional spikes that represent hot topics. By the time one is down to the 154th row this drops to about 7.5 bits. An example spike was at theme C(=C(C(Cl)(Cl)Cl)C*l)(Cl)Cl ranked 176th with about 7.5 bits was a popular one with 22 companies. The distribution continues to zero, although surprises of 1 bit or more are reported. Negatives can occur when companies avoid common chemical themes, but that best shows when using alert measure $\Delta I(A;B;C;...) = I(A;B;C;...t_2) - I(A;B;C;...t_1)$which is essentially the difference between Equation.1 at a particular period of time relative to preceding times; then a company may be more clearly revealed to have dropped its interest in a chemical theme. Typically many companies may be involved in 3 related chemical themes and a large number of single companies are associated with themes satisfying the query with associations worth 1-2 bits.

Of subsequent interest has been the use of the values (numbers of bits) of these associations to predict what existing chemical themes and companies a query chemical theme would be most likely be associated with, and with what strength (information). A very simple study consists of removing an existing chemical theme from the data current collection of them and finding which chemical themes it most closely matches. This has been done for a variety of types of compound, but particularly those of the steroid-like class of interest to us [17]. Details of this will be given elsewhere, but briefly, recall that one obtains a list of chemical themes ranked in terms of their association with one or more companies, any row in the output being a distinct theme and associated companies. The important and typical finding (but not obviously expected finding) is that associations of the same and similar value tend to imply very similar chemical themes. They may be from the same compound, but not always. In consequence, any chemical theme in a row removed, and then used as a query as if it were a potential new composition of matter will strongly tend to detect related themes that lie adjacent or near to the row from which it was removed.

## Discussion and Conclusions

The number of information bits for issues mentioned above, plus a few other relevant numbers that place constraints on possible values, are given in Table 1. Note that if we imagine a future case of a claim that considers a 1000 atom molecule of 50 possible types of atom at any location, the claim could certainly be made for a specific example embodiment and of course it may well be synthesizable in principle because that has nothing to do with the fact that the number of other possibilities exceeds the number of atoms in the observable universe. By Equation.2 the information content of that claim would be no more than for a typical narrow claim today. On the other hand if it made a claim for more than $10^5$ chemical themes, by the same Equation.2, the information content of the claim would approach 0 bits. Of particular interest in the discussion has been the idea that the information content by the above definitions is of the approximately same order as the information content required for a degree of selective molecular recognition. In many respects the examiner and assignee are in the position of using a digital molecular library, the distinction being that similarity is being selected in the search case, and complementarities in the experimental case. Complementarities as physical molecular recognition at a target nonetheless remain of interest if we are to discourage the assignee from deliberately or inadvertently covering

| System | Estimate of Information in Bits |
|---|---|
| $Log_2$ of number of 1000-atom molecules with all possible combination of 50 bonding atoms of B,C,N,O,F columns of periodic table | 5643 |
| $Log_2$ of number of atoms in observable universe $10^{78}$-$10^{82}$ | ~266 |
| $Log_2$ of number of atoms in/on planet Earth (6 x $10^{28}$) | 93 |
| $Log_2$ of number of valid chemical theme formats | 76 |
| $Log_2$ of number of potentially feasible valid chemical themes | ~56 |
| $Log_2$ of limiting number of molecules that might be produced and detected by future nanotechnological manufacturing | 56 |
| Information in DNA on Earth that distinguishes all known species | ~50 |
| $Log_2$ of number of CAS-registered organic, inorganic compounds, and biopolymer sequences | 29 |
| $Log_2$ of number of known and CAS-registered organic and inorganic compounds [45] | 26 |
| Strong molecular recognition interaction | 23-25 |
| $Log_2$ of number of patents in the world | 24 |
| Specified molecule with respect to example embodiments and specific compounds in US patents (circa 2008) | 23 |
| $Log_2$ of number of number of combinatorial chemistry generated compounds per annum in a company | 17 |
| Information content a DNA-protein binding site | 13-20 |
| $Log_2$ of number of variations for significant biological Action of a chemical theme non-polar, hydrogen bond donor, hydrogen bond acceptor, positively charged, | 11-19 |
| Claim of one chemical theme | 16-17 |
| Overlapping but context free chemical theme | 15 |
| Fair Markush claim | 13 |
| Estimated lower limit for significant molecular recognition and hence biological action | 13 |
| Information content (Equation 2) of claim containing 10 chemical themes | 13 |
| Information content (Equation 2) of claim containing 100 chemical themes | 10 |
| Association of a chemical theme with one or more companies | 10 |
| Strong association constant between two chemical themes | 9 |
| Typical association of a specific query of a class of molecule with one or more companies | 7-8 |
| Information content (Equation 2) of claim containing $10^4$ chemical themes | 3.5 |
| A priori information content of a chemistry claim based on claim allowance rate | ~1 |
| Information content (Equation 2) of claim containing large number (>>$10^5$) of chemical themes | ~0 |
| A priori information content of a chemistry claim based on the refutation principle | 0 |

**Table 1:** Information Contents (bits) of Interest in Thinking about Novel Compositions of Matter.

a vast number of different applications. This is also possible to do digitally, at least for the researcher to show supporting evidence of due diligence (Table 1).

The original study [17] included not only finding similar compounds on the patent data base, but used other criteria, notably satisfying computer simulations of binding at a pre-specified target.

## References

1. Adams RS (2006) Information Sources in Patents. Walter de Gruyter, Amsterdam, Netherlands.
2. Lynch MF, Barnard JM, Welford SM (1981) Computer Storage And Retrieval Of Generic Chemical Structures in Patents. J Chem Inf Comp Sci 21: 151-161.
3. Downs GM, Barnard JM (1998) Chemical Patents and Structural Information: The Sheffield Research in Context. J Doc 54: 106-120
4. Oldach S, Stabinsky N (2008) The Value of Patent Analytics. Intellectual Property Today.
5. Feldman R, Sanger J (2006) The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press, New York, USA.
6. Berks AH (2001) Current State of the Art of Markush Topological Search Systems. World Patent Information 23: 5-13.
7. Li J, Robson B (2000) Bioinformatics and Computational Chemistry in molecular Design: Recent Advances and their Application 101: 285-307. Peptide and Protein Drug Analysis Marcel Dekker.Inc, New York, USA.
8. Paolini GV, Shapland RH, van Hoorn WP, Mason JS, Hopkins AL (2006) Global Mapping Pharmacological Space. Nat Biotechnol, 24: 805-815.
9. Chen YP, Chen F (2008) Identifying targets for drug discovery using bioinformatics. Expert Opin. Ther Targets. 12: 383-389.
10. Patent Structure Handling in Torus
11. Waikato Internet Firm Creates Hub Of World Class Technology.Netvalue
12. Tripos Mol2 File Format
13. Scientific Enterprise Software. Accelrys
14. http://en.wikipedia.org/wiki/Molecule_mining #cite_note-kti03-0
15. Rhodes J, Boyer S, Kreulen J, Chen Y, Ordonez P (2007) Mining Patents using Molecular Similarity Search. Pac Symp Biocomput 304-315.
16. Engelfriet A (2006) Determining the Scope of a Patent. Ius Mentis: Law and Technology Explained.
17. Robson B, Dettinger R, Peters A, Boyer SK (2011) Drug discovery using very large numbers of patents: general strategy with extensive use of match and edit operation. J Comput Aided Mol Des 25: 427-441
18. Chen Y, Spangler S, Kreulen J, Boyer S, Thomas D et al. (2009) SIMPLE: A Strategic Information Mining Platform For IP Excellence. IBM Research.
19. Weininger D (1988) SMILES, A chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comp Sci 28: 31-36.
20. Regular Expressions (1997) The Single UNIX ® Specification, The Open Group.
21. http://perldoc.perl.org/perlre.html
22. Fisanick W (1990) The Chemical Abstracts Service Generic Chemical (Markush) Structure Storage and Retrieval Capability. 1. Basic Concepts. J Chem Inf Comp Sci 30: 145-154.
23. Barnard JM (1991) A Comparison of Different Approaches To Markush Structure Handling. J Chem Inf and Comp Sci 31: 64-68.
24. Barnard JM (1993) Substructure searching methods: old and new. J Chem Inf Comp Sci 33: 532-538.

25. Barnard JM, Downs GM (1997) Chemical Fragment Generation and Clustering Software. J Chem Inf Comp Sci 37: 141-142 .

26. Downs GM, Barnard JM (1997) Techniques for Generating Descriptive Fingerprints in Combinatorial Libraries. J Chem Inf Comp Sci 37: 59-61.

27. Barnard JM, Downs GM (1992) Clustering of Chemical Structures on The Basis of Two-Dimensional Similarity Measures. J Chem Inf Comp Sci 32: 644–649.

28. Brown RD, Martin YC (1996) Use Of Structure-Activity Data To Compare Structure-Based Clustering Methods And Descriptors For Use In Compound Selection. J Chem Inf Comp Sci 36: 572-584.

29. Robson B, Finn PW (1983) Rational Design of Conformationally flexible drugs. ATLA-Altern Lab Anim 11: 67-78.

30. Ivanciuc O (2008) Canonical numbering and constitutional symmetry. Handbook of Chemoinformatics, Wiley-VCH.

31. Daylight Chemical Information Systems, Inc.

32. Dethlefsen W, Lynch MF, Gillet VJ, Downs GM, Holliday JD (1991) Computer storage and retrieval of generic chemical structures in patents. 12. Principles of search operations involving parameter lists: matching-relations, user-defined match levels, and transition from the reduced graph search to the refined search. J Chem Inf Comp Sci 31: 253-260.

33. Franzosi M (2003) Markush Claims in Europe. Associated Lawyers Franzosi Dal Negro Setti, Milan, Rome.

34. Nakagama H, Heinrich G, Pelletier J, Housman DE (1995) Sequence and structural requirements for high-affinity DNA binding by the WT1 gene product**.** Mol Cell Biol 15: 1489-1498.

35. Robson B, McBurney R (2012) The role of information, bioinformatics and genomics: Drug Discovery And Development: Technology In Transition. (2nd Edn), Elsevier Press, Amsterdam, Netherland.

36. Robson B (1974) Analysis of the Code Relating Sequence to Conformation in Globular Proteins: Theory and Application of Expected Information. Biochem J 141: 853-867.

37. Robson B (2003) Clinical and Pharmacogenomic Data Mining: 1. The generalized theory of expected information and application to the development of tools. J Proteome Res 2: 283-302.

38. Robson B, Mushlin R (2004) Clinical and Pharmacogenomic Data Mining. 2. A Simple Method for the Combination of Information from Associations and Multivariances to Facilitate Analysis, Decision and Design in Clinical Research and Practice. J Proteome Res 3: 697-711.

39. Robson B (2004) The Dragon on the Gold: Myths and Realities for Data Mining in Biotechnology using Digital and Molecular Libraries. J Proteome Res 3: 1113-1119.

40. Robson B, Vaithiligam A (2010**)** Drug Gold and Data Dragons: Myths and Realities of Data Mining in the Pharmaceutical Industry. Pharmaceutical Data Mining, John Wiley & Sons, USA.

41. Robson B (2005) Clinical and Pharmacogenomic Data Mining: 3. Zeta Theory As a General Tactic for Clinical Bioinformatics. J Proteome Res 4: 445-455.

42. Robson B (2008) Clinical and Pharmacogenomic Data Mining: 4. The FANO Program and Command Set as an Example of Tools for Biomedical Discovery and Evidence Based Medicine. J Proteome Res 7: 3922-3947.

43. Popper K (1959) The Logic of Scientific Discovery. Routledge Classics, London and New York.

44. Probabilistic directed acyclic graphical model. Bayesian network

45. CAS registry number