J Comput Aided Mol Des (2011) 25:427–441
DOI 10.1007/s10822-011-9429-x

Drug discovery using very large numbers of patents. General strategy with extensive use of match and edit operations

Barry Robson · Jin Li · Richard Dettinger · Amanda Peters · Stephen K. Boyer

Received: 7 December 2010 / Accepted: 12 April 2011 / Published online: 3 May 2011

Abstract A patent data base of 6.7 million compounds generated by a very high performance computer (Blue Gene) requires new techniques for exploitation when extensive use of chemical similarity is involved. Such exploitation includes the taxonomic classification of chemical themes, and data mining to assess mutual information between themes and companies. Importantly, we also launch candidates that evolve by "natural selection" as failure of partial match against the patent data base and their ability to bind to the protein target appropriately, by simulation on Blue Gene. An unusual feature of our method is that algorithms and workflows rely on dynamic interaction between match-and-edit instructions, which in practice are regular expressions. Similarity testing by these uses SMILES strings and, less frequently, graph or connectivity representations. Examining how this performs in high throughput, we note 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.

Keywords Patent analytics · Drug discovery · Ligand design · Regular expression · Similarity

Introduction

Background

As an aid to drug discovery, we explore novel techniques for rapid analysis and use of large chemical data bases, including those obtained by mining patents, combined with protein target binding calculations. Information about patents and protein targets provide rich information for the pharmaceutical industry [1–9], management of which has become an industry in itself (e.g. Refs. [10–13]). But many applications of such data require use of the notion of *similarity* between molecules (e.g. Refs. [14–17]), often a rate limiting step in workflows [15]. It was so in our use of a patent data base generated by using Blue Gene to read automatically all US patents [16, 17]. The 6.7 million records comprise SMILES code [18] for compounds mentioned[1] along with assignee, and also the patent reference by which other data can be joined.

This report also includes work done in part by authors Barry Robson, Amanda Peters and Stephen K. Boyer at IBM Corporation.

B. Robson (✉)
St Matthews University School of Medicine, Grand Cayman, Cayman Islands, The University of Wisconsin-Stout, Menomonie, Wisconsin, USA
e-mail: brobson@smu.ky

J. Li
Global Compound Sciences, AstraZeneca R&D, Alderley Park, Macclesfield, Cheshire, UK

R. Dettinger
Prentice, Rochester, USA

A. Peters
Department of Physics, Harvard University, Cambridge, MA, USA

S. K. Boyer
Collabra Inc., San Jose, CA, USA

[1] All US patents are read, and chemistries other than pharmaceutical are thus extracted. These are retained for reasons of detecting prior art in compositions of matter, relevance to chemical taxonomy, repurposing for pharmaceutical applications, and studies relating to toxicity.

Discovery from patent data

Whilst it is evident that methods for rapid analysis of such data can provide important chemical business intelligence (which we also explore), it is less evident how they can lead *directly* to novelty. But a compound *not* similar in whole or part to any on the data base, and screened *in silco* to bind in an appropriate manner at the protein target, is a promising candidate for further study. We automatically evolve molecules, or more precisely *queries* about molecules that can evolve also in exactness of match. They are chosen from the patent data base, or simply from molecules specified as interesting lead compounds. They evolve with an implied fitness function, though in practice a candidate must not be similar to one on the data base in order to be considered for binding calculation. Other data bases of commercially available compounds are also used with the patent data base to support novelty, stability, and feasibility as further fitness criteria, as described briefly. This includes quantum mechanical calculations, and confirmation that chemical mini-fragments ("groups") are known to be tractable chemistry, because they occur in the data bases. In addition, as a check on binding calculations, extensive studies were performed to detect possible extensive binding site accommodation by the target (exemplified in the "Appendix").

Familiar and novel features of this work

There are many ways of looking at molecular similarity (e.g. Refs. [21–31]). We use *regular expressions* ("regex") [19] in the Perl environment [20] in order to manage both SMILES and graph (connectivity) representation. The method is flexible, but here come close to the LINGO method [14, 15] that compares SMILES strings by *fragmentation*: in LINGO, benzene c1 ccccc1 in SMILES becomes "c1cc, ccc1, 1ccc, cccc, cccc". However, the main difference is that the regex interact not only with SMILES but with each other in a particular *dynamical* way, rather like heterogeneous (i.e. non-identical) automata. In the present report, we apply *six basic principles* of collective behavior of match-and-edit objects such as regex. Based on these, the program ChemBiz comprises algorithms and associated workflows that take a few seconds or minutes to organize and use information from the 6.7 million compounds on a standard personal computer. The above principles are exemplified by building taxonomies of compounds, data mining the chemical business space, and generation of potentially novel candidates.

Theory

Dynamic systems of match-and-edit instructions

It suffices to consider the general match-edit object /A/B/m−n, p supported by a very few additional flow and conditional action features. The A in /A/B/m−n, p is at least part of another such object that must be matched. B is what the part matched, or another part referred to in A, may be changed to, with probability p. In practice, A and B are linear strings of characters. However, this includes also connectivity or graph representations in linear string format and in practice as sets of several such match-edit objects applied collectively. Here, parameters m−n become particularly important. They signify a range such as 1–3, meaning A must be found one to three times within an object for a match to occur. 0–0 means there is a match if there is no match by A. 0–5 means that there is no match if A if found more than five times, and 7-? means that A must match at least seven times. All the properties or capabilities of the objects required here are within the power of the regex s/A/B/with our extensions m−n and p. The effect of m−n can be encoded in /A/ in a regex, but these parameters play such a key role in our algorithms that they are placed "up front".

Because we emphasize interactions of these match-edit objects as constituting a potentially complex dynamical system, it is helpful to think of biological analogies such as a protein interaction network. An enzyme can recognize and modify parts of other enzymes with probability p (for a population of enzymes, seen as rate), so altering their role in the network. Similarly the A in /A/B/m−n, p can be part of a object /C/D/m′-n′, p′, and B can be part of an object /E/F/m″-n″, p″ that is generated by the action of /A/B/m−n, p on the object containing A. Recognition by enzymes can vary in specificity and sensitivity, and analogously the parts /A/ and /B/ in /A/B/m−n, p can contain not only the content A and B such as chemistry, but also contains additional *reserved features* that manage matching and editing, especially partial matching which allows variations in A to still be considered a match. ChemBiz ensures the legality of chemical content, and of the reserved features that in regex are *metachacharacters* [19]. An arbitrary collection or *soup* of such objects can have interesting ongoing and emergent properties. However, in some analogy with biological evolution of networks, the present report adds structure to the modes of interaction. We here use six principles of primarily axiomatic character as working guidelines. The first three relate in a familiar way, but in the language of the above objects, to graphs as the underlying representations of molecules. The second three relate to the forms imposed on the interaction network as a

graph. Note that in considering below a set $\{X_{m-n}\}$ of members X_{m-n}, not only X but also m−n may be different for each member.

1. Composition. If nodes in a given set of distinct graphs can be assigned names X so that no two given graphs are the same in composition as described by such names, then there exists a set {/X/Z/m−n, p} that collectively uniquely matches one such given graph.
   - Such a composition test using the set {/X/Z/m−n, p} may be the sequential application against graph of composition $\{X_{m-n}\}$ of separate pure match tests /A/m−n, /B/m′-n′, /C/m″-n″,…equivalent to /A/A/m−n, 0 /B/B/m′-n′, 0 /C/m″-n″, 0… with exit from the tests on the first non-match, and the order is immaterial except for efficiency.
   - To guarantee uniqueness of graphs in a previously unforeseen set of graphs, unique names X such that $m = n = 1$ are required; however we are often interested in cases where the composition of node names is not unique in a graph and an appropriately small number of matches are returned by appropriate choice of all the m−n.
   - To approach the case of uniqueness of graphs by composition of node names and so approach unique matches, node names as strings of characters can always be extended (further qualified and differentiated) beyond their initial given names.
   - When node name extension depends on features that do not specifically include the node names of other atoms, e.g. on features such as chirality that relate to the simplex representation of the graph (i.e. as n nodes maximally displayed in $n - 1$ dimensions), or that the node is in a cyclic path of node connections or a path of particular number of connections, then in matching a graph to a larger graph as a subgraph of that larger graph, the node name extension in the smaller graph can always be the same as that in the larger graph.
   - Node name extension can also use information from other node names from that local region of the graph where the named node resides, or equivalently from connections (arcs, edges) between nodes including the named node. However, when matching a graph to a larger graph as a subgraph of that larger graph, then the node names in the extension of node names in the subgraph are at least a sub-collection (sub-composition) of those in the larger graph matched.
   - Nonetheless, if node name extension includes nodes names, sets of graphs can exist in which such information matches directly because it describes features of connectivity that are the same throughout those graphs.
2. Exact Graph Match. If a given set of distinct graphs can be assigned connection names X so that no two given graphs are the same in composition described by such names, then there exists a set {/X/Z/m−n, p} that collectively uniquely matches one such given graph. A connection name may be $[X] = X_0, X_1, X_2, X_3,\ldots,X_n$ as a path in the graph; that is, a connection name is a list of node names of nodes in which contiguous node names are connected. Then there exists a set {/[X]/Z/m−n, p} that matches exactly one graph, and by sufficient extension of node names, a set {/[X]/Z/1-1, p} that matches exactly one graph.
   - If there are b nodes with node names $X_0, X_1, X_2, X_3,\ldots,X_b$ then such a match is a special case of a node name composition match with $X_1, X_2, X_3,\ldots,X_b$ as the node name extension of node named X_0, but allowing that a node may have several extended names or equivalently be seen as one name based on a set of such extended names. The meaningful distinction is that a node name extended by connections is a breadth first and a connection name is a depth first description of connected nodes. Thus the points of Principle 1 apply, including that it can be tested in the same way.
   - The tested set of names can in principle comprise a mix of node names and connection names.
   - If there are b nodes with node names $X_0, X_1, X_2, X_3,\ldots,X_b$ considered in a connection name, then b can be 0,1,2,3,…not exceeding the number of nodes in the graph that it relates to, then increasing b increases the power of discrimination. The value of b need not be the same for each connection name, but if all possible paths up the length b are represented, we can always discard connection names extended by less than b node names without losing information.
   - If in $X_0, X_1, X_2, X_3,\ldots,X_b$ the X_0, and X_b are the same node, either because of direction of construction or a cyclic path, then one connection name is redundant, and a canonical choice (i.e. preceding in alphanumeric order) should be made for the one retained.
   - There also always exists a set of unique connections called W,[X] where X does not use not unique node names and W is information based on the rest of the graph. However, the same information residing in W can always be provided by extending at least one duplicated node name.

3. Subgraph Match. If the set {[X]} of the smaller graph is a subset of the corresponding set for the larger graph being matched, and if there are several distinct graphs being matched and no two graphs have the same composition of names of connections, then there exists a set of {/[X]/Z/m-?, p} that uniquely matches at least one subgraph of those distinct graphs.
   - Points from Principles 1 and 2 apply.
4. Hierarchic Action of Matching and Editing. An arbitrary set of /X/Z/m−n, p acting on each other is not guaranteed to be a generative grammar.[2] However, without generating badly-formed objects a tree graph of permissible interactions representing a hierarchy of action may be traversed convergently or divergently.[3]
   - Content X and Z may nonetheless represent a general graph.
   - In principle, for such a hierarchy of action each object generated need not have the same grammar for the content and the same reserved features for managing partial matching, because it can be redefined to relate to each hierarchic level.
5. Evolution. If a member of a set of X in the set /X/Z/1,1, p where /X/ may also contain partial match features is transformed by that object into an object of general format /A/B/1-1, p with probability $0 < p < 1$, then that object is the sole descendent of X if the transformation successfully occurs. The sequence of /A/B/1-1, p so generated from X by repetition of the transformation is an evolution of that member X to the last /A/B/1-1, p as descendant. A set of the transforming objects of general format /X/Y/1-1, p always exists that implies for any X the probability density description of all possible evolutions of X, and a subset of more than one of these objects may exist that implies for any X and final /A/B/1-1, p the probability density description of all possible evolutions of X to that to the last /A/B/m−n, p as descendant.
   - In practice, each object generated has content with the same format and grammar of reserved match-edit features, because each object in the evolving sequence is to be tested against a large set of X in the same way.
   - However, at each step the reserved match-edit features objects generated may describe a scope (degree of generality) for matching that may be greater or less than that of the object from which it was derived by the action of a transforming object /X/Z/1,1, p.
6. Taxonomy. If two or more members of a set of X that map to the Z subset /X/Z/1-1,1, where /X/ contains partial match features, are transformed by that object into one object of general format /X_1/B/1-1,1 similarly containing partial match features that match those two or more members of the set of X, then that new object implies a set to which the two or more members belong. If two or more such second generation objects generated of general format /A_2/B/1-1,1 are similarly transformed into a further third generation object that matches those two more second generation objects, and so on, then the sequence of /A/B/1-1, p so generated from the two or more X is a taxonomy of those X as a tree graph of relationships between the X.[4]
   - Each object generated at a different taxonomic level need not have the same grammar for content and reserved match-edit features, because in practice this need not be used for matching but to name a set of objects as taxonomic level descriptors.
   - The relationship between these names and the implied format and grammar of the reserved match-edit features is nonetheless precisely defined by the objects /X/Z/1-1,1 that created them by matching and editing.

Note that while in ChemBiz A and B in /A/B.m−n, p are in practice always strings, the set and connectivity or graph data types can also be represented by use of metacharacters, or equivalently by separating the regex into several and using simple appropriate programming techniques (first bullet point of Principle 1). The regex data types so implied are for efficiency. They are not each closed universes because of further regex that interconvert them or sets of them. However, inter-conversion is almost always more

[2] Such a grammar is a set of continuously applied replacement rules that allows creation from a start object of any permissible grammatical object but no other objects in the language. The grammar includes that of the reserved features for managing the match and edit of content. It is only possible to disprove a sound generative grammar by doing the computation and halting at a badly formed object.

[3] Strictly speaking, this is an assertion. In the present chemical context, we have not found a counterexample that is not a foolish choice or "bug". It is complex because the grammar of an object and its reserved match-edit features may be correct, but the legality of content depends on perception of reasonable chemistry. Not least because some objects are input, we check automatically at intervals that the object grammar is legal, and that X is a valid chemistry at very least with permissible valancies.

[4] In other words for this and what follows, the outcome is that if several objects are matched by one object, and several of the latter kind of objects are matched by one object, and so on, it implies a taxonomy, but it is not a requirement the reserved match-edit features have the same format and grammar at each level.

time consuming, and is minimized by making extensive use of the first two of the following regex.

SMILES regex

Here A and B in /A/B/m−n, p are SMILES as strings of characters but also typically contain regex metacharacters. Hydrogen atoms are not expressed. Chemical entities of less than ten characters (e.g. small counterions) are ignored in the present study, and the largest of separate components in entries is always taken. SMILES chirality markers are retained. Also retained are the "metacharacters" of SMILES' own grammar (such as brackets that relate to branching structure), which frequently also happen to represent regex metacharacters. In order that SMILES metacharacters are not to be seen as regex metacharacters, they are specifically flagged ("literalized"), except for SMILES in the original patent data base. Many workers would consider any extensive use of SMILES regex to be doomed to failure, or at very least a big step backwards from graph methods, SMARTS representations, formula motif fingerprints, and a battery of multiple and hybrid methods chemical similarity methods (e.g. Refs. [10–15, 20–25, 30, 31]). The problem is that same molecule or part of such can be written as many different valid SMILES representations. Canonicalization [29] eliminates the problem for whole compound matches, but partial matches are widely considered troublesome. Small chemical changes sometimes alter the entire canonical representation. In the language of diagnosis and prediction, methods that allow for this are more *accurate* because they have both good *sensitivity* (fraction that is true from the positives) and good *specificity* (fraction that is true from the negatives) to the implied chemistry. The SMILES regex has excessive *specificity* and so may fail to find a similar chemistry. We refer to all related issues as *SMILES variance*. More accurate methods are used to test and analyze them, and generally "clean up" afterwards, appropriate at least with the present use cases.

Compositional regex

However, SMILES regex can also be used with an *extended notion of composition* (Principle 1) so that they are not at all susceptible to SMILES variance. In contrast to SMILES regex, they have considerable sensitivity, but a sufficient set can be quite specific to chemical structure. For example, *20Q Queries* input by a user are named after the game Twenty Questions, because issuing a *20Q query set* can pin down one or few molecules in a similar manner. An example 20Q query set is *Cl, 1; C, 30–38; =O, 2–4; #C, 1* (triple bond); *N, 0; S, 0–2; chirals, 4-?* (chiral centers); *branches 10–20*; *rings, 5–6; double overlapping rings 3–4*, converted to SMILES regex internally. Note the extension to the normal idea of chemical composition that resides in the convenient input token words such as *chirals*, *rings*, and *branches.* These features are readily derived from SMILES and SMILES regex, and a 20Q query set can be automatically extracted from a SMILES, promising probe, or genus (chemical theme, described below). Forms like *C=O, 2–5* work if in the canonical SMILES) we do not see O=C. A 20Q query set can be used to construct a single more complex 20Q query. For example the tokens *amino-acid-like, nucleotide-like, lipid-like, and sugar-like* followed by m−n will match any molecule containing m−n times a description covering these commonly occurring corresponding chemistries in biology, albeit it not exclusively. A 20Q query set can, in a special *search* command, be combined with query logic to other query information such as assignee organization, years or year range of publication, any other information to which the patent reference points, and the state of the system.

Graph or Connectivity Regex

These can have perfect accuracy in the limit of sufficient information content added, seen as sufficient extension of the notion of composition. Regex discussed here do not each carry the content of an entire graph, but rather each carries just one of a set of composition descriptions and the corresponding set of regex overall essentially makes up a 20Q query set. Each regex does not in general carry a fragment of a SMILES string: parsing of SMILES is required to interpret SMILES brackets in terms of an actual or implied connectivity matrix (e.g. Refs. [21–27, 31]). Double and triple bonds are not distinguished. Since the matrix is sparse, the concept of a limited set of named interactions occurring, expressible in the manner of a 20Q query, makes good sense. The format in each regex can vary. In the *atom* or *node name* method (Principle 1), the names involved are the essentially the normal atom names with extensions called *suffixes* (here in lower case solely for clarity of reading) describing local connectivity. For example, $C_{OS,C1,C1O}$ is the extended name of first carbon in OCSC(C)O, with an added touch of distinguishing power in that C1 in indicates that it is bonded to just one other C. This is a *breadth first* representation, based on *atom, directly bonded neighbor atoms, second order neighbor atoms* … etc., here up to three bonds away, when it is said to be a b = 3 representation. This approach requires use of regex metacharacters to handle differences in atom names due to "missing atoms" in a fragment when compared with the larger molecule matched. In contrast, *connection names* (Principles 2, 3) are depth first. We could emphasize the relationship to the extended atom name method by writing each atom with a suffix which is the set of all paths of

length b from it in alphabetic order, but it is better to write (for b = 3), a connection name simply as e.g. BrCCO (equivalent to OCCBr, the former alphabetic order being chosen as canonical). This is because there is then no effect of "missing atoms": the connection names that would have missing atoms are simply *missing* from the 20Q-style query set describing the fragment. They are more clearly related to LOGOs [14, 15]. Only paths b long are counted, and each can occur m > 1 times. For an exact match m-m is used and for a fragment match m-?. In the course of the study we tried a number of representations. Formats exemplified by use of atom name $C_{OS,C1,C1O}$, a larger set of connection names such as CSCO, and mixed descriptions such as $C_{OS,C1}$—$O_{C,S}$ all include some information up to b = 3 and give similar results for overall statistics regarding whether any two non-canonical SMILES are chemically equivalent. Later in the project we used a compressed binary format[5] that still represent match-edit objects and obtain similar results for each choice of b, but more rapidly.

Uses of SMILES regex

SMILES regex dynamics dominate ChemBiz. Table 1 shows the hierarchic nature of the current system (Principle 4) for the SMILES regex component. Important features are expanded upon in the following Sections.

Probe evolution

A series of SMILES regex can be used to repeatedly and automatically re-edit a *probe* which is also a SMILES regex (Principle 5). Analogy with Darwinian evolution is helpful. The probe evolves by *random mutation*, surviving natural selection by *not* matching the data base (the *inner design cycle*) *and* by binding appropriately to a bimolecular target in computational chemistry calculations (the *outer design cycle*). What evolves is not a representation of a compound but a partial match query against compounds on the patent data base; a probe for most of its life is somwhat like the Markush representations [6, 21, 22] used in patents. SMILES regex called *mutation operators* or *mutators* edit a matched probe with a probability $0 < p < 1$. This is important not only for modification of chemistry but also because it can alter metacharacters and increase ("wax") or decrease ("wane") the scope[6] of the probe on a probabilistic basis. Mutators reside on a *design file*. It is a Perl script and examples are fairly lengthy but a "primer" example is available on request. Probe evolution may run as branching parallel processes in the background. A probe is always tested against the patent data base and an evolutionary trajectory is normally aborted if it persistently matches the data base in $n_{mutations} = 1{,}000$ mutations. At preset intervals it can be tested more rigorously for potential novelty by rendering SMILES regex as graph regex. For selection by binding to the biomolecular target, the probe must also no match the patent data base but be in, or converted to, a state that describes a specific molecule. In some cases, a specific compound arises too infrequently, in which case a forced waning called *freezing* is applied. A *simplest representative member* is generated which has fewest atoms.

Generation of chemical themes

See Principle 6. SMILES on the patent data base, and SMILES regex that unambiguously describe specific molecules, are a *species* by biological analogy. Typically starting from the patent data base, taxonomy generation is roughly like probe generation but always waxes in generality with probability $p = 1$. The first level of generalization above species is the *genus*. In this report it is synonymous with *chemical theme* although the latter can be redefined as another representation, say a graph regex. A molecule can contain one or more chemical themes, with the requirement that a chemical theme must be valid as a molecule by itself, with correct valencies when hydrogen atoms are assumed, and in principle capable of synthesis. An exception is made for certain types of monomer when detected as a theme repeatedly occurring in oligomers or polymers; that repeated theme is used.[7] Genera may be seen as analogous to substructures [23], fragments [24] and

[5] *Connectivity strings* are not regex but the fast and compressed form of these for exact binary matching methods, because as b increases the number of paths increases explosively. They are re-definable in the system and several implementations and variations are being explored. Currently, rows 2–4 and columns 5–9 of the periodic table are numbered 0–14, and iodine added as 15. These are then expressed in binary. The propyl fragment C–C–C–C is thus 0001000100010001, held in two 8-bit bytes or characters in memory.

[6] However, this is not very efficient because if a probe is matching the patent data base, so will a waxed form of it, i.e. one broader in scope. Waxing should occur when there are no matches for a set number of mutations (editing iterations). Waning should occur both (a) in the case of matches against the patent data base and (b) at intervals when there persistently *no* matches in order to generate specific compounds. A critical parameter is thus a default or user-specified number of mutations which have elapsed before a probe is deemed "persistently novel" and worthy of waned to a specific candidate compound. Nonetheless, the optimal choice of parameter can vary from study to study, and it is best to start to wane when the novelty entropy (see above Footnote) reaches a critical value. .

[7] In one mode, this information and that from small complete molecules can be used to select the chemical themes in a first pass, and not overlap their chemistries when they are seen in larger compounds. However building blocks can be made out of building blocks, and we prefer an empirical approach based on association information described below.

Table 1 Hierarchy of SMILES regex

Action order	Objects	Probe evolution	Taxonomy evolution	Comments
4	Control operator	SMILES regex that queries and edits the Generalization-Level Operator to optimize production of candidate ligands from the probe.	SMILES regex that queries and edits the Generalization-Level Operator to optimize production of candidate ligands from the probe.	More typically, may also query and edit a Mutation Operator directly. Control Operators are created or edited manually in the report, but in a prototype they can be modified automatically by so-called higher level *Executive Operators*.
3	Generalization-level operator	SMILES regex that queries and edits the Mutation Operator to ensure that the probe becomes less specific ("waxing"), or more specific ("waning")	**SMILES regex that queries and edits the Mutation Operator to ensure that the Generic Representation will be less specific, so generating a Generic Representation for the next taxonomic level up**, or more specific to attempt to deduce a more specific Generic Representation.	Probabilistic, but may be manually determined via a Control operator. **Once useful taxonomies are found, the Generalization-Level Operators are deterministic and are held on a Perl file called the *taxonomy file*, generating a higher taxonomic level as needed.**
2	Mutation operator acting on internal data object.	**SMILES regex that queries and edits the probe**.	SMILES regex that queries and edits the Generic Representation	**Probabilistic.** **In the case of the probe a set of Mutation Operators are applied sequentially and they are stored on an external Perl file called the *design file*.**
1	Internal data object derived from input.	***Probe* as a SMILES regex**	***Generic Representation* as a SMILES regex**	May also be a specific compound represented by a SMILES regex literalized and with no metacharacters
0	Input data object matched by internal data object	**A SMILES entry on patent data base identified by a query, or specified separately as input by user.**	**All SMILES entries on patent data base.**	Not literalized, i.e. all symbols are interpreted as SMILES characters, not regular expression metacharacters

Functionalities most important to the present report are highlighted in bold

sometimes fingerprints [25]. In this study the requested limiting fragment size for the genus is 10–35 SMILES characters (without such a restriction, the representation of a molecule is a *generic representation*, not a genus). The numbers of characters in the genus can be more or (typically) less than requested size because they are "repaired" to describe a synthesizable molecule or monomeric unit in its own right, at least in principle. The ability to depart from a single grammar of content and of reserved match-edit features is exploited (see Principle 6). At each taxonomic level, the definition should seek to capture the description of molecules somewhat in chemist's terms, and give a reasonable distribution of members between the groups so classified. Currently, the first level of generalization as genus (chemical theme) would convert SMILES SCCCCSiCC=O to SC*=O which is read as "all molecules with sulfur attached to a chain of two or more carbon atoms, with a carbonyl oxygen at the other end of the chain". In detail, the process for generating genera from SMILES on the patent data base is equivalent to following recipe for generating genera.

1. Chemical entities such as inorganic or organic ions which are not chemically bonded are removed. In case of doubt of which is the main part, the larger is taken.
2. Non-required characters @ (chirality)/(cis–trans), square brackets [], ion charges and any associated charge numbers are removed. [CH2] etc. is considered joined to the main molecule. Hydrogen atoms if present are removed. The complete original SMILES code will appear in output catalogs, and thus the original form where stated will have the chiral decription etc.
3. SMILES strings of 35 characters or less of Canonical SMILES code are further processed without fragmentation.
4. In the usual fragment taxonomy option, overlapping strings of 35 characters are extracted by starting at

the first character of the SMILES code of the patented example molecule and stepped to the right one character step at a time, i.e. there is a frame shift employed of one character. Note that these strings as preliminary fragments can grow or shrink in processing with reference to original SMILES representation, this being part of a curation process as follows, which results in the fragments proper.

5. All fragments which do not start with an atom, i.e. characters such as C or Br, are discarded. All fragments with unclosed right brackets ')' are discarded. Fragments with unclosed left brackets '('are simply healed by addition of the appropriate number of brackets to the right of the fragment.
6. Connection numbers are renumbered so that the first is 1. Monomeric units, fragments of rings, etc. as incomplete structures are allowed in the present study, so connection numbers which point out of the chemical theme and have no internal partner are retained, and renumbered relative to that assigned 1. In the present study, the ChemBiz option used was that connection numbers which emerge from the above renumbering as less than 1 or greater than 9 are deleted. Alternatively, < and > may be used to indicate the points of deletion respectively.
7. Except for C, O, S, N, and P, elements, i.e. atoms types, are replaced by a representative member of their column in the periodic table (which is generally the element most familiar to the pharmaceutical industry). In the columns of the periodic table in which the above C, O, S, N, and P, another "second best" representative is selected, e.g. Si stands for all other elements in the column of the periodic table in which carbon occurs. So 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 the actual associated regex the alternative elements (atom types) are specifically listed in a regular expression logical 'OR' construct.
8. A run of 2 or more elements say in general X without intervening symbols is replaced by X*.
9. This is extended for carbon C. A run of two or more carbons C without intervening brackets or triple bonds (# in SMILES) is replaced by C*. But there may be several single and double bonds (= in SMILES) present which are not preserved.
10. Solely hydrocarbon side chains without connection numbers and not starting with a double or triple bond are replaced by (C), even if there are brackets and hence hydrocarbon branches implied within it. In *any* side chain runs which *start* with double or triple bonds, it is retained as in (=C*). In non-hydrocarbon side chains (i.e. with non-C atoms present), sub-themes within, such as C*(C)O, are retained, i.e. brackets are not removed, though (C) may be a branched hydrocarbon.
11. Optionally, after the above, to represent succinctly highly repetitive chain molecules, nested repeat patterns are expressed in brackets {}. Patterns may embed patterns, but once done, multiple runs of characters {{.. and}}.. are converted to single {and}.

Higher levels of taxonomic description

The *family* or *second generic level* is simply the genus in which all SMILES brackets are removed. In the *order* or *third generic level* all non-alphabetic characters are removed and the reverse order of characters is considered equivalent. These give a reasonable spread of members. The *class* or *empirical formula* puts atom types in alphabetic order and the number seen in the order representation is specified after each atom, e.g. BrC3N2O. It can be expressed as a 20Q query set. The *phylum* or *composition* lacks the numbers, implying 1-? for m−n in each 20Q query. The *kingdom* or *industry* is convenient in partitioning the data by industry interests: o = organic including carbon and oxygen, x = carbon but no oxygen, h = hydrocarbon (carbon and hydrogen only), and i = inorganic (no carbon).

Data mining chemistries

The patent data base can be data mined for *degree of association* between chemical themes, or of many assignees with chemical themes, or association between patent references or data to which patent references point. The present study uses the above genera based on SMILES. While this works for graph descriptions as chemical themes, it is efficient to apply graph based matches and resolve chemical equivalence later. There can be many themes and many companies associated, so data is often sparse. The result of integrations over Bayes posterior distributions yields measures of surprise from both sparse and extensive data in terms a linear expression of Riemann zeta functions ζ, here sufficiently seen as the Euler series $\zeta(s, \quad n) = \Sigma_{k=\quad 1,2,3,\ldots,n} \quad k^{-s} = 1 + 2^{-s} + 3^{-s} + \ldots n^{-s}$ [32, 33]. In the present report, we use this for a form of Fano's mutual information [32] $I(A;B;C;\ldots) = \zeta(s = 1, o[A,B,C,..]) - \zeta(s = 1, e[A,B,C,..])$. Here o[] and e[] are observed and expected frequencies of events A, B, C,… For no data, the measure equals zero, and generally $e[A,B,C,\ldots] = N^{1-n}\ o[A]o[B]o[C]\ldots$ given total amount of data N. The concept of an *alert* is similar but is specifically reserved to mean *a change in surprise*, i.e. of I(A;B;C;…), say changes in the current year as opposed to

previous years. Similar measures based on ζ are also used in assessment of the entropy in taxonomic distributions and the openness of the search space in the vicinity of the probe, but are not discussed here.

Methods

Patent text analytics [16, 17] was applied to all US patents to extract chemical formulae [34] using the Blue Gene very high performance computer [35]. However, the resulting patent data base can reside on a personal computer, and all the use cases discussed below were covered in a software demonstration seminar of less than 1 h using only a standard laptop. That is with the exception of quantum mechanical, docking, and molecular dynamics studies, also done on Blue Gene and described below. The "Appendix", and methods related to Ref. [36] to be described elsewhere, attempt to predict this binding by personal computer. We employed IBM's middleware DDQB (Data Discovery and Query Builder) [37] on the laptop, to provide the working ChemBiz environment and interface. DDQB is an abstraction layer which takes large heterogeneous data sources. It can interpret complex ChemBiz queries against such by perceiving data as objects with ontological relationships, i.e. semantically. Via DDQB we accessed the Internet (for Web-based research) and the ZINC data base [38] and Protein Data Bank [39]. The system can also interconvert with Mol2 format [12], e.g. to pass to other applications such as visualization.

Results

Initial research

Although now historical, an example earlier e-mail alert that did ultimately trigger the present study was "Licorice Root May Keep Mental Skills Sharp" [40]. This described the action of licorice in reducing glucocortocoid levels in the brain, and linked ultimately (e.g. Ref. [41]) to natural steroid-like inhibitory compounds with an anti-inflammatory action as, in particular, 11-beta-hydroxysteroid dehydrogenase type 1 inhibitors. These could be described by the following 20Q queries: *C, 30–38; = O, 2–4; O, 6–8; N, 0; rings,5–6; chirals,4-?* which matched 0.02% of records on the patent data base, reported by ChemBiz as a catalog of companies showing the patented compounds that match. Companies such as Hoffman-la-Roche, Dong Kook and Finetech were well represented, but Biorex, a UK company, was prominent with 44 compounds.

Launching the probe

The probe was launched to run in the background while further studies below proceeded. The command *maximum tries* by default selects the first of the records matched by the 20Q described above, as initial state of the probe. *Maximum tries = 1,00,000, mutations = 500* evolves the probe 500 mutations before restarting from the record with 20Q matches described above. It continues until 100,000 mutations occur overall. *Maximum hits* (i.e. matches) specifies how many records match the current probe state before the sweep through the patent data base is abandoned.

Taxonomy generation

Command *catalog: themes* searched all 6.7 million records and extracted 833,333 genera of primarily 9–14 atoms. It appears impossible to obtain equally populated meaningful genera (see Sect. 4.4); the distribution approximately follows Zipf's law [42] (essentially, the ζ function of Sect. 2.9) but with spikes (i.e. a multimodal component). The average overall yield of genera is 0.12 per record, though recall that they are based on SMILES regex and so many of these are chemically equivalent. By generating from genera the equivalent of various graph connectivity representations with b = 3 (Sect. 2.4), in 13% of genera groups of up to 3 were chemically equivalent, in 2% 4–6, and in 75% 7–28. This was surprising as, a priori, a smaller fragment such as COC(S)(C)O could be written in non-canonical SMILES some 40 reasonable ways. Increasing b on samples of these does not significantly change this distribution, except that some 2% now had groups of 50 or more that were chemically equivalent. The reason seems to be a mix of factors. Many entries are not extensively asymmetric or chiral, and many are approximately of fragment size. Very few of all alternative theoretically possible chemical themes are on the patent data base and appear to be at least partially constrained by the use of canonical SMILES; many chemically equivalent genera seen as fragments in data base entries appear to have different surrounding chemistries: see Discussion and Conclusions. There appear to be recurrent biochemical and synthetic themes. Moreover, very often the same genera appeared to associate with certain companies, explored as follows.

Data mining all chemical themes from all patent records

The command *catalog: rank by association, themes, companies* generated (in 15–25 min according to personal computer) a spreadsheet in which single chemical themes are ranked by association strength. All associated assignees

are shown alongside each, along with a count of how many times they specified in patents distinct molecules with that theme. The distribution of number of uses of the chemical themes by companies in a row decreases down the list. This again follows Zipf's law with occasional spikes that now more clearly represent hot topics in the pharmaceutical business. Relatives of purine appear to dominate the higher association strengths. The top of the list at 684 centinats (hundredths of natural logarithmic unit) is N(C*1)C(C*(N=2)C(C) C=NC2N3CNC*3 associated with Berlex and 33 other companies. This means that they referenced to this theme 934 times more than would be expected on a chance basis from abundance of each company and the genus in the data base. Recall that in generic representation '*' means one or more C's, and connections out of the fragment as for (C*1) above are retained, but renumbered from 1. Shortly after in the list at 650 centinats C(CN(C(C*(N=1)C)=NC1N2CNC*2)C*N provides an internally complete fragment. By the time one is down to between the 154th rows (519 centinats) down to the 175th (154 centinats), there are only about one to three companies per theme, but an example spike is the chlorocarbon theme C(=C(C(Cl)(Cl)Cl)C*1)(Cl)Cl ranked 176th at 519 centinats, popular with 22 companies. If two genera have the same association in centinats then they are similar. With a difference of greater than 10 centinats the chemistries are typically unrelated. Similar genera can occur on up to 30–40 consecutive rows, typically with several assignees in common. On average a genus in this kind of output has about 7–8 chemically equivalent other genera. Other with 13–14 had low degrees of association, and others with near zero and negative association are not listed. These are in some sense "less meaningful chemical themes".[8] For *similar* genera there is roughly an equal mix of types of difference: genera distinguished only by SMILES variance shows approximately the same frequency of having similar association strength as do minor chemical modifications and overlapping themes.

Data mining focused by 20Q queries

This repeats Sect. 4.4, but now with the patent data base filtered by the same 20Q queries as in Sect. 4.1, to focus on steroid-like compounds. The records so extracted represent 270 chemical themes (genera) as generated only from those

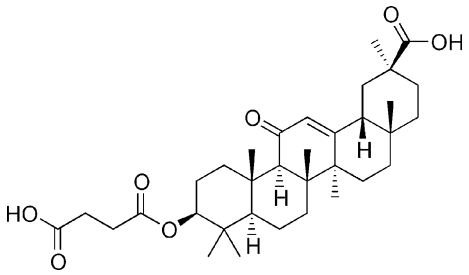

Fig. 1 Carbenoxolone

records. They are 0.02% of all the data base records that yielded 833,333 genera, so 167 genera might have been expected. As in Sect. 4.1, Biorex emerged prominently, but not alone, occurring just 13 times in the top 41 rows of genera. Internet research (e.g. Refs. [43, 44]) showed that the analogue commercially released by Biorex was carbenoxolone (Fig. 1). The enzyme 11-beta-hydroxysteroid dehydrogenase type 1 (which converts cortisol to cortisone) showed up as the most cited target for carbenoxolone on our directory of ligand targets,[9] and in querying against the Protein Data Bank yielded 1HDC.pdb and 2BEL.pdb. 2BEL became the candidate binding target, being a well refined structure conveniently including bound carbenoxolone inhibitor. Carbenoxolone was found to be the commercial product of Biorex. Although it is an analogue of the licorice root extract glycyrrhizic acid, the latter is substantially modified in vivo prior to binding to target. Thus it is not studied here, and we used carbenoxolone as our reference point in binding studies.

Feasibility screening of preliminary candidates

Although workers in the field use empirical directories of unstable groups and perhaps deliberately avoid electronically complex molecules that require QM methods to explore, we chose to use extensive minimization of Hartree–Fock calculations with GAMESS [45] parallelized on Blue Gene. This not only tests for stability in the sense of maintaining the structure but also allows less common groups to be parameterized for binding studies, such as certain thioketones. Internal stability is not the only issue.

Thioketones showed up quite often in probe evolution because no such steroid-like derivatives in the probe matched the patent data base, e.g. those with thioketone C=S derivatives at C11 (steroid ring nomenclature). This could signal synthetic difficulty, or lack of experience, toxicity, or instability due to intermolecular intercations. Indeed, thioketones are typically reactive [49] and may oligermerize.

[8] Obviously, this idea requires further analysis. While they typically appear to be regions of overlap between recurrent biochemical and synthetic themes, this may reflect the likelihood that companies associated with them tend to use the same synthetic strategy, while different companies have a bigger chance of using different strategies. Prominent in strong associations are ring systems, possibly suggesting more obvious synthetic strategies, as well as more restricted Canonical SMILES solutions.

[9] But this was confirmatory rather than discovery because the connection showed up in Internet research in Sect. 4.1 (e.g. Ref. [42]), so we do not describe our methods of building directories here.

However, the command *search = AZ or ASTRA or ZENECA or IMPERIAL CHEM for =S* found thioketones of comparable (but non-steroidal) local chemistry along with thioxoacetamides, thioureas and carbothioyls. Other data bases most importantly the ZINC data base (as a collection of commercially available compounds) was also checked for both feasibility and potential novelty of promising candidates.

Ligand binding calculations

Ligand–protein binding calculations were done by DOCK with AMBER; the standard approach was taken except for parallelization on Blue Gene [46–48]. The NADP cofactors were retained. Binding of those candidates of Table 2 that have the steroid ring core is essentially like that for carbenoxolone (Fig. 1) which we take as our comparison point. Results for most successful candidates as judged by binding are shown ranked with increasing binding stability downward in columns 1 and 2 of Table 2, and those of particular interest are drawn in Fig. 2. Candidates cboNRing, cboS1, and cboS2 are predicted as binding as stably as carbenoxolone. There are differences negligible within the state of the art, because it is hard to agree with experiment to this level even within a related series of compounds. However, prediction of DOCK-AMBER results by another theoretical method on a standard machine seems tractable, and has been used to study binding of some candidates in more detail (see Column 3 and Appendix).

Discussion and conclusions

The technology really addressed here is not high performance computing, nor the regular expression, and is only in part about evolving candidates. We wish to convey two general messages. The first is that any appropriate match-and-edit instructions interacting in a dynamical way can be a valuable means of constructing algorithms and relevant workflows. The second is subtle because it concerns the question of when best to apply similarity tests of different accuracy, from the *final utility* perspective. It becomes more evident that chemical similarity and novelty are human concepts and context dependent, and can only be objectively considered in the context of a particular task.

In particular, we sometimes wondered whether a rejection as detected prior art was too stringent in *any* part of workflow for generating candidates, as if rigorous similarity tests between parts of molecules might be counterproductive to their fine intentions. Sometimes, fragments not matching because of SMILES variance could represent *appropriate misses*. Apart from wishing to avoid difficult syntheses due to certain surrounding chemistries, that for example C=S appears in a compound on the data base, while a fragment tested against it includes only the corresponding C, raises the issue of how well a subgraph match truly describes equivalent chemistry. Of course, reliance on SMILES variance to signal an appropriate non-match in such a specific case is an unreliable and arbitrary method; it only makes some sense as an overall statistical trend in high throughput. Within the available options of the current system, it would be better see the above carbon as an entity with an extended name, say $C_{C,C=S}$ (in this case describing directly bonded atoms and bond order), and demand that C_{CC} does not match. Better still, the use of regex metacharacters in the suffix could cover a variety of acceptable and unacceptable match cases. Unfortunately, extension of such thinking to a knowledge based solution, one that addresses *many* aspects of overall utility, would require the outcomes of many real pharmaceutical projects in order to formulate quantitative rules of match. It is too early for our approach to have this kind of data.

Nonetheless, the patent data base itself represents historical data about what the industry thought was a novel and useful piece of chemistry. Whilst the usual answer to "How many types of chemistry are there?" is "As many as you like" and evidently correct if we *arbitrarily* adjust our taxonomy schema, the picture changes when association information is also considered. The notion of a boundary of a piece of chemistry in a molecule appears to be inherent in association: in preliminary analysis some themes that overlap by three quarters of their atoms or more can have radically different degrees of association with companies. This along with results of Sects. 4.3 and 4.4, balanced against the considerations of the above paragraph, suggests that there are less than 100,000 distinct chemical themes on the patent data base that are meaningful and interesting to the industry. The growth of the list is often by the sudden appearance of the spikes of association; these will be interesting to investigate in detail. All this relates to important chemical business intelligence. However, sole use of such a historical perspective could hamper innovation. It seems best that an innovative medicinal chemist corrects, selects, and extends the final ranked candidate list prior to synthesis and testing. As long as chemical informatics is perceived as having value, it seems reasonable that a large and growing ranked list of preliminary candidates can be of benefit in that way.

Appendix: Binding studies

Binding studies will be more familiar, but are well known to be nontrivial. Note first that in the catalytic region of chain A of the experimental 2BEL tetramer structure with

Table 2 Docking and preliminary stability data for compounds discussed in the text

Molecule	DOCK estimated free energy of binding Kcal/Mole (Blue Gene, AMBER force field)	Estimated free energy of binding: Robson-Platt force field (adjusted: see "Appendix")
Corphos	−16.8	−16.5
(cboNRing)* (3*β*)-3-[(3-carboxypropanoyl)oxy]-11-oxoolean-12-aza-30-oic acid	−16.4	−17.0
Carbenoxolone	−16.3	(−16.3)
(cboS2)* (3*β*)-3-[(3-carboxypropanthioxoyl)oxy]-11-oxoolean-12-en-30-oic acid	−16.3*	−16.8
(cboS1/cboS)* (3*β*)-3-[(3-carboxypropanoyl)oxy]-11-thioxo-olean-12-en-30-oic acid	−16.3*	−14.3
(cboShort)* (3*β*)-3- [(propanoyl)oxy]-11-oxoolean-12-en-30-oic acid	−15.2*	−14.6
(cboVeryShort)* (3*β*)-3- [(ethanoyl)oxy]-11-oxoolean-12-en-30-oic acid	−15.1*	−14.9
Cortisone	−14.7	−14.7
Cortexolone	−14.6	−15.3
(cboS3VeryShort2)* (3*β*)-3-[thioacetyl]-11-oxoolean-12-en-30-oic acid	−14.5*	−14.7
(cboS3VeryShort1)* (3*β*)-3-[thioacetyl]-11-oxoolean-12-en-30-oic acid carboxy enantiomer	−14.4*	−13.7
Azanaphthalene derivative 1: (2,3)-dithio-6-hydroxy-8-carboxy-(1,7,9)-azanaphthalene	**	−14.3
Hydrocortisone	−14.3	−14.1
Predisone	−14.1	−10.1
Steroid hyperoxide	−13.4	−10.8
ZINC00004088 (S')-5-ami-7(ethoxycarbonylamino)-2-methyl-3-phenyl-1,2-dihydropyrdo[3,4-*b]* pyrazin-6-ium	−11.2	−10.5
ZINC00001964 ethyl hydroxy4-7,8-demethoxypyramido[4-5-*b*]-quinolone-2-carboxylate	−11.2	−13.3
ZINC00006561 N-(3-chloro-1,4-dioxo-1,4-dihydronapthalen-2-yl)-2-(2-fluorophenyl)acetamide	−10.7	−9.1
Azanaphthalene derivative 2: (2,3)-diethyl-6-hydroxy-8-carboxy-(1,7,9)-azanaphthalene	**	−9.4
ZINC00027264 (RE)-4-(2-(7-chloro-5,8-dioxo-708-dihydroquinolin-6(5H),ylideneamino)ethyl)morpholin-4-ium	−10.6	−9.7
ZINC00001703 (5S,8S,9S,10S,13S,14S,17R)-1,10,-13-trimethyl-3-oxo-4,5,6,7,8,9,10,11,12,13,14,15,16,17–tetradecahydro-3*H*- cylopenta]phenanthren-17-yl acetate	−10.6	−9.6
ZINC00001328 8-chloro-2-methyl-5-(2-6(methylpyradin-3-yl)ethyl)-2,3,4,5-tetrahydro -1*H*-pyrido[4,3,-*b*]indol-2-ium	−10.6	−11.2
ZINC00023388 (*R*)-2-(3-methoxyphenyl)-6-(pyrrolidin-1-yl)-2,3-dihydroquinolin-4(1*H)*-one	−10.5	−9.8
ZINC00011032 (1R)-16,17-dimethoxy-11-azatetracyclo[9.8.03,8.014,19] nonadeca-3,5,7,14,(19),15,17-hexaen-11-ium (protonated N)	−10.5	−11.0

The results using the Robson-Platt force field were adjusted as described in "Appendix", except DOCK-AMBER results marked with an *asterisk**, which were predicted after that adjustment

carbenoxolone and NADP cofactor ligands (and rather similarly for the other monomers), the serine Ser 170 side chain oxygen atom makes a surprisingly tight 2.5 Å approach to the original double bonded C=O oxygen O11 atom on C11 of the steroid-like framework. The sum of the oxygen atom van der Waal's radii of the serine and ligand oxygen is 1.52 + 1.52 = 3.04 Å. Following DOCK and AMBER, this O…O distance is 3.0 Å. Adding to the electronegative tension, there is also a close approach by the phenolic oxygen of Tyr 183 at 5.0 Å, and by oxygen O7 N in the NADP cofactor at 4.9 Å. The other close approaches do not compensate significantly: they are most importantly the carbons C4 N of NADP at 4.3 Å and the Ala 172 side chain carbon at 4.5 Å, but a good binding comes from interactions involving the whole binding site region. It is substitutions at or around C11 that are compensated by the most surprising degree of binding site accommodation, i.e. significant conformational changes in backbone as well as side-chains in the binding site. Replacing the O…O contact by O…S by using thioketone (cboS1) as ligand initially generates a considerable van der Waal's *and* electrostatic repulsion as the thioketone S is

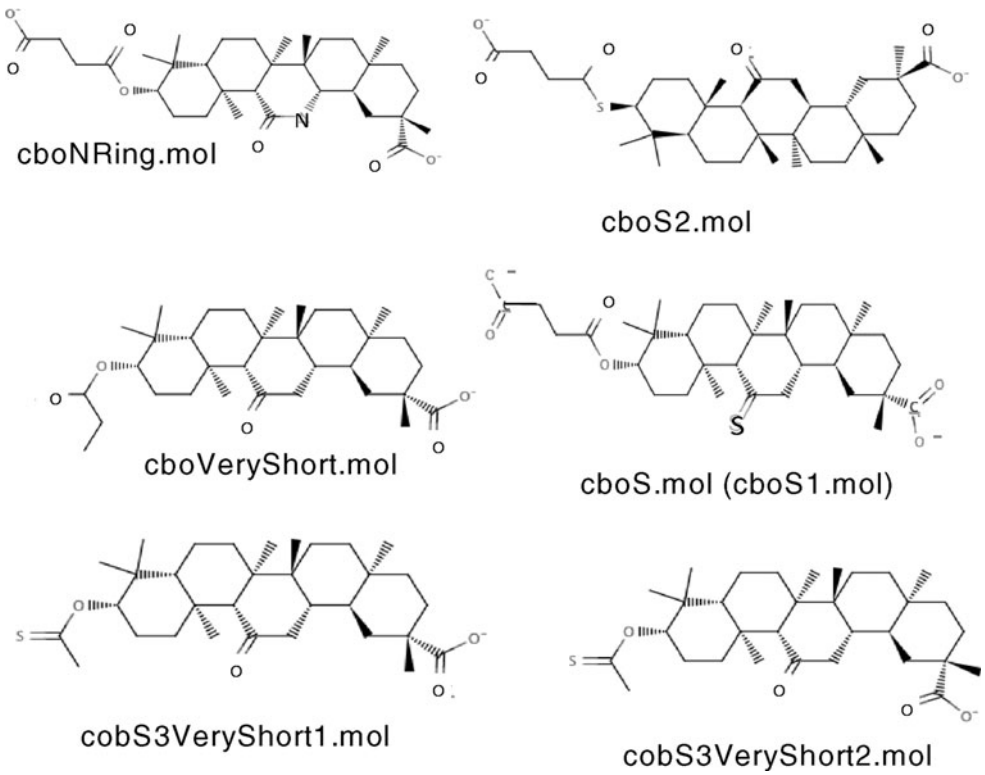

Fig. 2 Initial steroid-like designs passing novelty, stability, and binding criteria

strongly electronegative [49], but binding is still accommodated. These observations led to the following studies.

In order (a) to see if any multiple binding modes and binding site conformational changes may have been missed, and (b) to try and predict Blue Gene performance as a screening tool for potential candidates, a variety of extensive analyses and comparative studies using the full tetramer with NADP cofactors were performed using KRUNCH [50]. This can use and compare AMBER and other force fields and re-parameterize them, as well as perform both molecular mechanics and dynamics calculations. As an alternative to Blue Gene's processing speed and brute force calculation, this has a large kit of techniques for spanning energy barriers and searching conformational space, developed and adapted over many years by Robson and coworkers. See Ref. [51] for a general review, and techniques of Refs. [52–54] were also used. Results relative to carbenoxolone are shown in column 3 of Table 2. The regression slope of columns 2 and 3 is close to unity at 0.97 with a Pearson's correlation R of 0.88, with intercept close to zero. That is after adjusting slope for predictive purposes: prior to the latter, on average the RPFF energies were amplified 4.6% over the DOCK-AMBER results. The major difference here appears to be in electrostatics, because a simple 6% increase in effective dielectric constant in the RPFF brings the results into alignment (regression slope 0.95). Even recalling that this is predicting another calculation, not experiment, the above is a reasonable result within the state of the art. However, in more recent studies, probes started from the patent and ZINC data bases were left to evolve several weeks. Some ligands bound specifically to the 2BEL binding site by the RPFF plus heuristic methods, but *did not bind* to the site using DOCK and AMBER. An early example was (2,3)-dithio-6-hydroxy-8-carboxy-(1,7,9)-azanaphthalene as indicated by '**' in the second column of Table 2. The common feature of these "prediction failures" is that they are small ligands of one or two rings, and display multiple binding modes.

Note added in proof It may be worth highlighting that the bulk of this report is primarily theoretical and presents our strategy, and adding comment on how we see the road ahead. We also provided a brief but fairly broad scope of example applications and results, and they already raise to our minds many interesting questions and opportunities. In particular, results regarding the usefulness of SMILES regex compared with connectivity methods are necessarily preliminary, not least because an inevitable arbitrariness in the empirical assignment of genera as chemical themes affects the distributions of such descriptions that can be considered chemically similar. Yet enthusiasm for an simple extensive quantitative analysis taking account of variation in such assignment is somewhat blunted by the broader issue of what even equivalence of fragment chemistry can mean in practice (see "Discussions and conclusions"). In the way forward, a global rather than local, piecemeal perspective of the issues addressed needs to be retained, since those issues are not independent. Our comments regarding what really matters, the amount of information in the relationships between chemical themes, their definitions, and with laboratories, applications and general utility,

relate to the broader information "phase" space of chemistry. It is this as a whole that we see as the primary front begging exploration.

References

1. Adams RS (2006) Information sources in patents. Walter de Gruyter: Amsterdam, The Netherlands
2. Lynch MF, Barnard JM, Welford SM (1981) Computer Storage and retrieval of generic chemical structures in patents, 1. Introduction and general strategy. J Chem Inf Comp Sci 21(3): 148–150
3. Downs GM, Barnard JM (1998) Chemical patents and structural information: The Sheffield research in context. J Documentation 54(1):106–120
4. Oldach S, Stabinsk N (2009) The value of patent analytics, 2008. Intellectual property today. http://www.iptoday.com/articles/2008-6-oldach.asp. Accessed 20 Mar 2009
5. Feldman R, Sanger J (2006) The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge University Press, Cambridge
6. Berks AH (2001) Current state of the art of Markush topological search systems. World Patent Inf 23(1):5–13
7. Li J, Robson B (2000) Bioinformatics and computational chemistry in molecular design. Recent advances and their application. In Peptide and Protein Drug Analysis, Marcel Dekker NY, 285–307
8. Paolini GV, Shapland HBR, van Hoorn WP, Mason JS, Hopkins AL (2006) Global mapping pharmaceutical space. Nat Biotechnol 24(7):805–815
9. Chen YP, Chen F (2008) Identifying targets for drug discovery using bioinformatics. Expert Opin Ther Targets 12(4):383–389
10. Digital Chemistry (2009) Digital chemistry. http://www.digitalchemistry.co.uk/prod_torus_patent.htm. Accessed 20 Jul 2009
11. Reel Two, Reel Two web site (2007) http://www.reeltwo.com/. Accessed 20 Jul 2009
12. Tripos Inc (2008) http://www.tripos.com/data/support/mol2.pdf. Accessed 5 Apr 09
13. Symyx, Symyx Web Page (2009) http://www.symyx.com. Accessed 10 Nov 2009
14. Grant JA, Haigh JA, Pickup BT, Nicholls A, Sayle RA (2006) Lingos, finite state machines, and fast similarity searching. J Chem Inf Model 46(5):1912–1918
15. Haque IS, Pande VS, Walters WP (2010) SIML: A fast SIMD algorithm for calculating LINGO chemical similarities on GPUs. J Chem Inf Model 50:560–564
16. Rhodes J, Boyer S, Kreulen J, Chen Y, Ordonez P (2007) Mining patents using molecular similarity search. pacific symposium on biocomputing, Maui, Hawaii, 3–7 January 2007 Ed. Altman et al. World Scientific Publishing; p 304–315, http://www.almaden.ibm.com/asr/projects/biw/publications/Rhodes.pdf
17. Chen Y, Spangler S, Kreulen J, Boyer SK (2009) SIMPLE: A strategic information mining platform for IP excellence. In: IEEE international conference on data mining workshops, Miami, Florida, 6 Dec 2009. p 270–275. http://domino.research.ibm.com/library/cyberdig.nsf/papers/95D73078344701C9852576350055DBF3/$File/rj10450.pdf
18. 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
19. The Open Group, Regular Expressions (2009) The Single UNIX ® Specification, Version 2, 1997. Opengroup.org. http://www.opengroup.org/onlinepubs/007908799/xbd/re.html. Accessed 1 Aug 2009
20. Wall L, The Perl Development Team (2006) Perl.org. http://perldoc.perl.org/perlre.html. Accessed 9/1/2009
21. Fisanick W (1990) The chemical abstracts service generic chemical (Markush) structure storage and retrieval capability. 1. Basic concepts. J Chem Inf Comp Sci 30(2):145–154
22. Barnard JM (1991) A comparison of different approaches to Markush structure handling. J Chem Inf Comp Sci 31(1):64–68
23. Barnard JM (1993) Substructure searching methods: old and new. J Chem Inf Comp Sci 33(4):532–538
24. Barnard JM, Downs GM (1997) Chemical fragment generation and clustering software. J Chem Inf Comp Sci 37(1):141–142
25. Downs GM, Barnard JM (1997) Techniques for generating descriptive fingerprints in combinatorial libraries. J Chem Inf Comp Sci 37(1):59–61
26. Barnard JM, Downs GM (1992) Clustering of chemical structures on the basis of two-dimensional similarity measure. J Chem Inf Comp Sci 32(6):644–649
27. 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
28. Robson B, Finn PW (1984) Rational design of conformationally flexible drugs. ATLA Journal. Alternatives to Laboratory Animals 11: 67–78
29. Ivanciuc O (2003) Canonical numbering and constitutional symmetry. In: Handbook of Chemoinformatics, Ed. J. Gasteiger, Wiley-VCH, pp 139–160
30. Daylight Chemical Systems, Inc (2009) http://www.daylight.com/. Accessed 10 Apr 2009
31. Dethlefsen W, Lynch MF, Gillet VJ, Downs GM, Holliday JD, Barnard JM (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(2):253–260
32. 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
33. 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(9):3922–3947
34. Wikepedia (2010) http://en.wikipedia.org/wiki/IUPAC_nomenclature. Accessed 8/30/2010
35. Wikepedia (2010) Wikepedia. http://en.wikipedia.org/wiki/Blue_Gene. Accessed 8/3/09
36. Kramer A, Horn HW, Rice J (2003) Fast 3D molecular superposition and similarity search in databases of flexible molecules. J Comp Aided Mol Des 17(1):13–38
37. IBM Corporation, Data Discovery and Query Builder's User's Guide (2006) IBM Corporation. http://publib.boulder.ibm.com/infocenter/systems/topic/ddqb/v2r1ddqbusersguide.pdf. Accessed 7 Apr 2009
38. University of California San Francisco, http://zinc.docking.org/. Accessed 8 Aug 2009
39. RCSB Protein data Bank (2008) http://www.wwpdb.org/docs.html. Accessed 5 Apr 2009
40. Warner J (2004) Licorice root may keep mental skills sharp: compound derived from licorice root may fight effects of aging on brain. 2004, March. WebMD News. http://www.webmd.com/alzheimers/news/20040329/licorice-root-may-keep-mental-skills-sharp. Accessed 5 Apr 2009
41. Livingstone DE, Walker BR (2003) Is 11beta-hydroxysteroid dehydrogenase type 1 a therapeutic target? Effects of carbenoxolone in lean and obese Zucker rats. J Pharmacol Exp Ther 305(1):167–172

42. Wikepedia (2009) http://en.wikipedia.org/wiki/Zipf's_law. Accessed 6 Aug 2009
43. CAS, a division of the American Chemical Society. Support Page (2009) http://www.cas.org/support/scifi/index.html. Accessed 1 Jan 2010
44. CAS, a division of the American Chemical Society, Products page (2009) http://www.cas.org/products/sfacad/index.html. Accessed 1 Jan 2010
45. Schmidt MW, Baldridge KK, Boatz JA, Elbert ST, Gordon MS, Jensen JH, Koseki S, Matsunaga N, Su S, Windus TL, Dupuis M, Montgomery JA (1993) General atomic and molecular electronic structure system. J Comp Chem 14:1347–1363
46. Peters A, Lundberg M, Sosa CP, Lang T (2007) High throughput computing validation for drug discovery using the DOCK program on a massively parallel system. 1st annual MSCBB, Northwestern University, Evanston, IL, September 2007; available as Peters A, Lundberg M, Lang T, and Sosa, CP, 2008, RedPaper 4410 from IBM Corporation Poughkeepsie, NY
47. Balius TE, Mukherjee S (2008) Stony Brook University web site. http://www.ams.sunysb.edu/~tbalius/NamdandDockonNYBlue.pdf. Accessed 8 Aug 2009
48. Shivakumar D (2008) (updated 2009). University of California San Francisco, http://dock.compbio.ucsf.edu/DOCK_6/tutorials/amber_score/amber_score.htm. Accessed 12 Aug 2009
49. McWeeny R (1979) Coulson's Valence, 3rd edn. Oxford University Press, Oxford, UK see Ch. 6
50. Robson B, Curioni A, Mordasini T (2002) Studies in the assessment of folding quality for protein modeling and structure prediction. J Proteome Res (Am Chem Soc) 1(2):115–133
51. Robson B, Vaithilingham A (2008) "Protein Folding Revisited" pp 161–202 in Progess in Molecular Biology and Translational Science, Vol 84: Molecular Biology of Protein Folding, Elsevier Press/Academic Press
52. Robson B, Douglas GM, Platt E (1982) A new algorithm for rapid calculation of conformational energies. Biochem Soc Trans 10:388–389
53. Robson B, Platt E (1986) Refined models for computer calculations in protein engineering. Calculation and testing of atomic potential functions compatible with more efficient calculations. 188: 259–281
54. Collura VP, Greaney PJ, Robson B (1994) A method for rapidly assessing and refining simple solvent treatments in molecular modeling. Example studies on the antigen-combining loop H2 from FAB fragment McPC603. Protein Eng 7:221–233