(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2020-35115
(P2020-35115A)
(43)公開日　令和2年3月5日(2020.3.5)

(51)Int.Cl.		F I		テーマコード（参考）
G06F 16/00	(2019.01)	G06F	17/30	170F

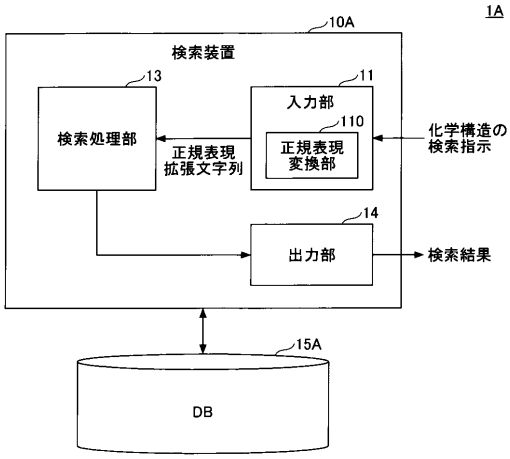
審査請求 未請求　請求項の数 11　ＯＬ　（全 17 頁）

(21)出願番号　特願2018-159778 (P2018-159778)
(22)出願日　平成30年8月28日 (2018.8.28)

(71)出願人　504203572
国立大学法人茨城大学
茨城県水戸市文京二丁目１番１号
(74)代理人　100107766
弁理士　伊東　忠重
(74)代理人　100070150
弁理士　伊東　忠彦
(72)発明者　藤芳　明生
茨城県日立市中成沢町四丁目１２番１号
国立大学法人茨城大学　工学部内

(54)【発明の名称】検索システム、検索方法、及び検索プログラム

(57)【要約】
【課題】簡単かつ柔軟に化学物質を検索することのできる構成と手法を提供する。
【解決手段】検索システムは、化学構造を表わす分子記述言語に正規表現を適用して拡張した表現形式の正規表現拡張文字列を取得する入力部と、前記正規表現拡張文字列をもとに、データベースから該当する化学物質を抽出する検索処理部と、前記検索処理部による検索結果を出力する出力部と、を有する。
【選択図】図１



【特許請求の範囲】
【請求項１】
化学構造を表わす分子記述言語に正規表現を適用して拡張した表現形式の正規表現拡張文字列を取得する入力部と、
前記正規表現拡張文字列をもとに、データベースから該当する化学物質を抽出する検索処理部と、
前記検索処理部による検索結果を出力する出力部と、
を有することを特徴とする検索システム。
【請求項２】
前記検索処理部の入力に接続されて、前記正規表現拡張文字列をグラフオートマトンに変換する変換部、
をさらに有し、
前記検索処理部は、前記グラフオートマトンにしたがって、前記グラフオートマトンで定義される状態遷移を満たす化学物質を前記データベースから抽出することを特徴とする請求項１に記載の検索システム。
【請求項３】
前記検索処理部は、前記正規表現拡張文字列を先頭から順に読み込み、前記正規表現拡張文字列を、前記化学構造の状態遷移を規定する前記グラフオートマトンに変換することを特徴とする請求項２に記載の検索システム。
【請求項４】
前記グラフオートマトンを保存する記憶部、
をさらに有することを特徴とする請求項２または３に記載の検索システム。
【請求項５】
前記入力部は、前記正規表現拡張文字列を入力として受け取り、前記正規表現拡張文字列を前記検索処理部へ渡すことを特徴とする請求項１に記載の検索システム。
【請求項６】
前記入力部は、既存の分子記述文字列または化学構造式を入力として受け取り、前記既存の分子記述文字列または前記化学構造式を前記正規表現拡張文字列に変換して、前記検索処理部に渡すことを特徴とする請求項１に記載の検索システム。
【請求項７】
前記検索システムは、ネットワークを介して接続される１以上の端末装置を含み、
前記端末装置は、前記入力部の機能の少なくとも一部と、前記出力部の機能の少なくとも一部を有することを特徴とする請求項１～６のいずれか１項に記載の検索システム。
【請求項８】
情報処理装置において、化学構造を表わす分子記述言語に正規表現を適用して拡張した表現形式の正規表現拡張文字列を取得し、
前記正規表現拡張文字列をもとに、データベースから該当する化学物質を抽出し、
抽出された結果を出力する
工程を含むことを特徴とする検索方法。
【請求項９】
取得された前記正規表現拡張文字列をグラフオートマトンに変換するステップ、
をさらに有し、
前記化学物質の抽出は、前記グラフオートマトンにしたがって、前記グラフオートマトンで定義される状態遷移を満たす化学物質を前記データベースから抽出することを特徴とする請求項８に記載の検索方法。
【請求項１０】
コンピュータに、
化学構造を表わす分子記述言語に正規表現を適用して拡張した表現形式の正規表現拡張文字列を取得する手順と、
前記正規表現拡張文字列をもとに、データベースから該当する化学物質を抽出する手順

と、
抽出された結果を出力する手順と、
を実行させる検索プログラム。
【請求項１１】
取得された前記正規表現拡張文字列をグラフオートマトンに変換する手順、
をさらに有し、
前記化学物質を抽出する手順は、前記グラフオートマトンにしたがって、前記グラフオートマトンで定義される状態遷移を満たす化学物質を前記データベースから抽出することを特徴とする請求項１０に記載の検索プログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、化学構造の検索システム、検索方法、及び検索プログラムに関する。
【背景技術】
【０００２】
分子の化学構造を文字列で表現する方法として、ＳＭＩＬＥＳ（Simplified Molecular Input Line Entry System）、ＳＭＡＲＴＳ（SMILES Arbitrary Target Specification）などの表現法が用いられている。ＳＭＩＬＥＳは、分子の化学構造をＡＳＣＩＩコートの英数字と記号で文字列化した表記法であり、構造検索やデータ入力などに広く用いられている。
【０００３】
ＳＭＡＲＴＳは、ＳＭＩＬＥＳを検索クエリに拡張した表記法である。ＳＭＡＲＴＳでは、結合の種類（二重結合または三重結合）や基の種類（芳香族または脂肪族）など、簡単な構造検索の条件を表記することができる（たとえば非特許文献１参照）。ＳＭＩＬＥＳやＳＭＡＲＴＳの外にも、ＩｎＣｈＩ（Information Chemical Identifier）など、いくつかの表記法がある。
【０００４】
一方、文書の検索において、正規表現が利用されている。文書の検索・照合において、文字列の正規表現から有限状態オートマトンを構築し、有限状態オートマトンを用いて文字列のマッチングを行う手法が提案されている（たとえば、特許文献１参照）。
【先行技術文献】
【特許文献】
【０００５】
【特許文献１】特許第３８５２７５７号
【非特許文献】
【０００６】
【非特許文献１】Daylight Theory Manual, Daylight version 4.9, Release Date 08/01/11, Daylight Chemical Information Systems, Inc., http://www.daylight.com/dayhtml/doc/theory/index.html
【発明の概要】
【発明が解決しようとする課題】
【０００７】
既存の分子記述言語による表現法では、検索したい化学物質の任意の集合を一つの文字列で表現することが困難である。任意の集合には、化学物質の部分構造、特定の性質を持つ分子の集合、反応により生成される生成物の集合などが含まれる。ＳＭＩＬＥＳで化学構造の集合を表現する場合、その集合に含まれる化学物質の文字列をすべて列挙しなければならない。列挙するには大きなスペースが必要であり、そもそも無限集合は表現できない。類似の構造を持つ化合物の集合を示す表記法としてマルクーシュ構造（Markush structure）が存在するが、化学構造の表現の制限が大きく、選択肢に含まれる置換基を列挙する自然語の説明文が、別途必要となる。

【０００８】
本発明は、簡単かつ柔軟に化学物質を検索することのできる構成と手法を提供することを目的する。
【課題を解決するための手段】
【０００９】
簡単かつ柔軟な化学構造の検索を実現するために、化学構造を表わす分子記述言語に正規表現を適用して拡張した正規表現拡張文字列を用いる。正規表現拡張文字列をもとに、データベースで該当する化学物質を検索する。
【００１０】
本発明の一態様では、検索システムは、
化学構造を表わす分子記述言語に正規表現を適用して拡張した表現形式の正規表現拡張文字列を取得する入力部と、
前記正規表現拡張文字列をもとに、データベースから該当する化学物質を抽出する検索処理部と、
前記検索処理部による検索結果を出力する出力部と、
を有する。
【発明の効果】
【００１１】
上記の構成により、簡単かつ柔軟に化学構造を検索することができ、検索範囲を拡張することができる。
【図面の簡単な説明】
【００１２】
【図１】実施形態の検索システムの一例を示す模式図である。
【図２】検索システムの別の例を示す模式図である。
【図３】検索システムの実現に適したハードウエア構成図である。
【図４】実施形態の検索システムの変形例を示す図である。
【図５Ａ】化学構造の正規表現拡張文字列を例示する図である。
【図５Ｂ】化学構造の正規表現拡張文字列を例示する図である。
【図６】グラフオートマトンへの変換を説明する図である。
【図７】グラフオートマトンに基づいて検索された検索結果の一例を示す図である。
【図８】検索画面の一例を示す図である。
【図９】「最小一致」の検索条件が選択されたときの検索結果の表示例を示す図である。
【図１０】グラフオートマトンへの変換の別の例を示す図である。
【図１１】データベースに格納されるテーブルの一例を示す図である。
【図１２】検索方法のフローチャートである。
【発明を実施するための形態】
【００１３】
実施形態では、化学構造を表わす分子記述言語に正規表現を適用して、多様な化学構造とその検索条件を一つの文字列で表現する。この明細書では、化学構造を表わす分子記述言語に正規表現を適用して拡張した表現形式の文字列を、「正規表現拡張文字列」と呼ぶ。分子構造の記述に正規表現を適用することで、化学構造における任意の集合（所定の構造の有無、任意の回数の繰り返し、置換基どうしの結合／分離など）や検索条件を、１つの文字列で記述することができる。
【００１４】
正規表現拡張文字列に基づいて化学物質を検索する際に、正規表現拡張文字列をグラフオートマトンに変換して、グラフオートマトンにしたがって検索してもよい。グラフオートマトンを用いることで、完全一致の他に、部分構造検索や、任意の集合の検索が容易になる。
【００１５】
グラフオートマトンは、グラフ理論を適用したオートマトンである。オートマトンは、

入力に対して、状態に応じた処理を行って次の状態に遷移させる仮想マシンであるが、ここでは、オートマトンにより規定された状態遷移の集合も「オートマトン」に含めることとする。

【００１６】

グラフ理論における「グラフ」は、点（頂点又はノード）の集合と、点から延びる辺（枝またはエッジ）の集合で表される構造である。化学物質データベースに保存されている個々の化学物質の構造は、グラフ理論におけるグラフとみなすことができる。データベースに保存されている既知の化学物質が、グラフオートマトンで規定された状態遷移に一致する構造を有するか否かを判定することで、検索が容易になる。

【００１７】

図１は、実施形態の検索システム１Ａの模式図である。検索システム１Ａは、検索装置１０Ａと、データベース（ＤＢ）１５Ａを含む。検索装置１０は、入力部１１、検索処理部１３、及び出力部１４を有する。

【００１８】

入力部１１は、化学構造の検索指示を入力として受け付ける。検索対象となる化学構造の入力形態は任意である。ユーザによって、直接、正規表現拡張文字列が入力される場合は、入力された文字列はそのまま検索処理部１３に入力される。検索対象が、既存の分子記述言語、構造式、マルクーシュ構造などで入力された場合は、正規表現変換部１１０で正規表現拡張文字列に変換されてから、検索処理部１３に入力される。

【００１９】

上述したように、正規表現拡張文字列は、分子記述言語に正規表現を適用して拡張された表現形式の文字列である。以下では、分子記述言語としてＳＭＩＬＥＳを例にとって説明するが、ＳＭＡＲＴやその他の分子記述言語に正規表現を適用してもよい。

【００２０】

ＳＭＩＬＥＳの主要な表記規則として、次のようなものがある。元素は元素記号で表示され、Ｃ，Ｎ，Ｏ，Ｐ，Ｓ，Ｃｌ，Ｂｒ，Ｉに結合する水素は記載されない。二重結合は「＝」、三重結合は「＃」、分岐は「（）」（小括弧）で表される。環の開始と終点となる原子に番号を付けるなどである。たとえば、プロパン（C_3H_8）は「ＣＣＣ」、イソブタン（C_4H_{12}）は「Ｃ（Ｃ）ＣＣ」、シクロヘキサンは（C_6H_{12}）は「Ｃ１ＣＣＣＣＣＣ１」と表記される、等である。

【００２１】

正規表現は、文字列の集合を単一の文字列で表現することができる。たとえば、「太郎または花子」は「太郎｜花子」と記述される。ここで縦棒「｜」は、選択肢を区切る表現である。大文字で始まり小文字が続く英単語は、[A-Z][a-z]+と記述される。プラス符号「＋」は、直前の表現が１個以上あることを示す。アスタリスク「＊」は、直前の表現が０個以上あることを示し、疑問符「？」は直前の表現が０個または１個あることを示す。携帯電話の電話番号は0[7-9]0-[0-9]{4}-[0-9]{4}と表現される。波括弧「｛｝」内の数字は、直前の要素の繰り返しの回数を示す。

【００２２】

ＳＭＩＬＥＳに正規表現を適用した文字列を「正規表現拡張ＳＭＩＬＥＳ」と呼ぶ。正規表現拡張ＳＭＩＬＥＳの定義の例をいくつか挙げる。

（１）すべてのＳＭＩＬＥＳの文字列は正規表現拡張ＳＭＩＬＥＳとすることができる。

（２）文字列w_1＝ａb_1ｃ，w_2＝ａb_2ｃ，…，w_n＝ａb_nｃを正規表現拡張ＳＭＩＬＥＳの文字列とする。この場合、ａ｛b_1|b_2|…|b_n｝ｃは、集合｛w_1，w_2，…，w_n｝を表わす正規表現拡張ＳＭＩＬＥＳである。

（３）文字列ｗ＝ａｂｃを正規表現拡張ＳＭＩＬＥＳの文字列とする。この場合、a{b}*cは、集合｛ａｃ，ａｂｃ，ａｂｂｃ，ａｂｂｂｃ，…｝を表わす正規表現拡張ＳＭＩＬＥＳである。

（４）文字列ｗ＝ａｂｃを正規表現拡張ＳＭＩＬＥＳの文字列とする。この場合、a{b}+cは、集合｛ａｂｃ，ａｂｂｃ，ａｂｂｂｃ，…｝を表わす正規表現拡張ＳＭＩＬＥＳであ

る。
（５）文字列ｗ＝ａｂｃを正規表現拡張ＳＭＩＬＥＳの文字列、ｉを整数値とする。この場合、a{b}_{i}cは、集合｛ａｂ^{i}ｃ｝を表わす正規表現拡張ＳＭＩＬＥＳである。
（６）文字列ｗ＝ａｂｃを正規表現拡張ＳＭＩＬＥＳの文字列、ｉ＜ｊを整数値とする。この場合、a{b}_{i,j}cは、集合｛ａｂ^{i}ｃ，ａｂ$^{i+1}$ｃ，…，ａｂ^{j}ｃ｝を表わす正規表現拡張ＳＭＩＬＥＳである。
【００２３】
ＳＭＡＲＴＳ等の他の分子記述表現を正規表現に拡張する場合も、同様に定義される。正規表現変換部１１０は、入力された分子記述表現に上記の規則を適用して、化学構造の正規表現拡張文字列を生成する。正規表現を適用することで、化学構造の中の特定の部分の繰り返し、特定の部位における選択肢、特定の構成要素の有無、置換基の結合または離脱などを、１つの文字列で記述することができる。さらに、繰り返し回数や選択肢の範囲を、無限、有限を含めて表現することができる。正規表現拡張文字列は、選択肢の範囲や、集合に含まれる要素数が多いほど効果的である。
【００２４】
標準ＳＭＩＬＥＳなどの既存の分子記述言語に替えて、化学構造式が検索対象として入力されたときは、正規表現変換部１１０は、化学構造を文字列化し、この文字列に正規表現を適用する。化学構造の文字列化は、たとえば、化学構造のあるひとつの頂点を選び、その頂点と辺で連結される隣接する頂点を順に選択して符号（原子記号を含む）を与えて文字列化する。環を形成しているところは、切り開いて、グラフ理論の「スパニングツリー」に変換する。このとき、環を切り開いたところにラベル付けをして、連結されていたもの同士を明示することで、文字列にすることができる。
【００２５】
検索処理部１３は、外部の化学物質データベースを参照し、また、必要に応じてデータベース１５Ａを参照して、該当する化学物質を抽出する。化学物質は、たとえばパターンマッチング等によって抽出される。パターンマッチングには最短一致、最長一致などが含まれてもよい。
【００２６】
出力部１４は、検索処理部１３による検索結果を出力する。検索結果は、正規表現拡張文字列を関連付けてデータベース１５Ａに保存されて、次回以降の検索に利用されてもよい。
【００２７】
図２は、実施形態の検索システム１Ｂの模式図である。検索システム１Ｂは、検索装置１０Ｂと、データベース（ＤＢ）１５Ｂを含む。検索装置１０Ｂは、入力部１１、変換部１２、検索処理部１３、及び出力部１４を有する。
【００２８】
入力部１１は、化学構造の検索指示を入力として受け付ける。検索対象となる化学構造の入力形態は任意である。ユーザによって、直接、正規表現拡張文字列が入力される場合は、入力された文字列はそのまま変換部１２に入力される。検索対象が、既存の分子記述言語、構造式、マルクーシュ構造などで入力された場合は、正規表現変換部１１０で正規表現拡張文字列に変換されてから、変換部１２に入力される。
【００２９】
変換部１２は、入力された正規表現拡張ＳＭＩＬＥＳの文字列を、グラフオートマトンに変換する。入力文字列からグラフオートマトンへの変換方法は、後述する。変換されたグラフオートマトンは、正規表現拡張文字列及び化学構造式と対応付けて、データベース１５Ｂに保存されてもよい。データベース１５Ｂに保存された情報は、次回以降の変換処理や検索処理に利用されてもよい。
【００３０】
検索処理部１３は、外部の化学物質データベースを参照し、また、必要に応じてデータベース１５Ｂを参照して、グラフオートマトンで定義される状態遷移を満たす化学物質を

抽出する。グラフオートマトンで定義される遷移状態を満たす化学物質は、正規表現拡張ＳＭＩＬＥＳで特定された集合に含まれる物質である。

【００３１】

検索は、たとえばマッチング判定によって行われ、外部のデータベースに格納されている化学物質の任意の頂点（ノード）から順に、グラフオートマトンで定義される状態遷移が満たされるどうかを判定していく。したがって、完全一致だけではなく、部分構造の一致も検索することができる。また、検索指示された集合の中の最小サイズの要素を検索する最小一致や、集合中の最大サイズの要素を検索する最大一致なども指定することができる。

【００３２】

出力部１４は、検索処理部１３による検索結果を出力する。

【００３３】

図３は、図１の検索システム１Ａ、及び／または図２の検索システム１Ｂの実現に適したハードウエア構成図である。検索システムは、たとえば、ネットワークに接続されたパーソナルコンピュータ（ＰＣ）１００によって実現可能である。ＰＣ１００は、ＣＰＵ（Central Processing Unit）１０１、主記憶装置１０２、補助記憶装置１０３、入力装置１０４、表示装置１０５、通信インターフェース（Ｉ／Ｆ）１０７、及びドライブ装置１０８を有し、これらの要素はバスＢによって相互に接続されている。

【００３４】

ＣＰＵ１０１は、主記憶装置１０２に格納されたプログラムに従って検索装置１０の動作を制御する。検索装置１０Ａ及び１０Ｂの検索処理部１３と、検索装置１０Ｂの変換部１２は、ＣＰＵ１０１によって実現可能である。

【００３５】

主記憶装置１０２には、ＲＡＭ（Random Access Memory）、ＲＯＭ（Read Only Memory）等が用いられ、ＣＰＵ１０１で実行されるプログラム、ＣＰＵ１０１による処理に必要なデータ、ＣＰＵ１０１の処理で得られたデータ等を記憶又は一時保存する。

【００３６】

補助記憶装置１０３には、ＳＳＤ（Solid State Drive）、ＨＤＤ（Hard Disk Drive）等が用いられ、各種の処理を実行するためのプログラム等のデータが格納される。補助記憶装置１０３に格納されているプログラムの一部を主記憶装置１０２にロードし、ロードされたプログラムをＣＰＵ１０１が実行することで、各種の処理が実現される。図２の変換部１２によって生成されたグラフオートマトンを保存するデータベース１５Ｂは、補助記憶装置１０３によって実現されてもよいし、外部のメモリを利用してもよい。

【００３７】

入力装置１０４は、マウス、キーボード等を有し、ユーザが検索装置１０で検索を行うときに必要な情報を入力する。表示装置１０５は、ＣＰＵ１０１の制御のもとに、入力画面、検索結果の出力画面などを含む各種の情報を表示する。ＰＣ１００がタブレットＰＣの場合、入力装置１０４と表示装置１０５が一体化されたタッチパネル式のディスプレイ（ユーザインターフェース）であってもよい。

【００３８】

通信Ｉ／Ｆ１０７は、ケーブル配線又は無線により、ネットワークを通じて通信を行う。たとえばＰＣ１００からネットワークを介して外部の化学物質データベースにアクセスして検索する場合、通信Ｉ／Ｆ１０７によって通信が行われる。

【００３９】

ドライブ装置１０８は、ドライブ装置１０８にセットされたＣＤ－ＲＯＭ（Compact Disc Read-Only Memory）等の記憶媒体１９と、ＰＣ１００との間のインターフェースをとる。

【００４０】

ＰＣ１００を検索装置１０として動作させるために、検索プログラムが用いられてもよい。検索プログラムは、ＣＤ－ＲＯＭ等の記憶媒体１０９によってＰＣ１００に提供され

てもよいし、通信Ｉ／Ｆ１０７を介してダウンロードされてもよい。ＰＣ１００にインストールされた検索プログラムは、ＣＰＵ１０１によって実行される。

【００４１】

プログラムを保存する記憶媒体１０９はＣＤ－ＲＯＭに限定されず、コンピュータで読み取り可能なデータの構造を有する一時的でない（non-transitory）有形の（tangible）媒体であればよい。ＣＤ－ＲＯＭの他に、ＤＶＤ（Digital Versatile Disk）、ＵＳＢメモリ等の可搬の記録媒体であってもよいし、フラッシュメモリ等の半導体メモリであってもよい。

【００４２】

図４は、実施形態の変形例としての検索システム１Ｃの模式図である。検索システム１Ｃは、サーバ装置２００と、ネットワーク２を介してサーバ装置２００に接続される１つ以上の端末装置３Ａ～３Ｎ（適宜、「端末装置３」と総称する）を含む。サーバ装置２００は、検索処理部２０と、データベース２５を有する。サーバ装置２００は、ネットワーク２を介して外部の化学物質データベース（ＤＢ）２６と接続されていてもよい。

【００４３】

各端末装置３は、図１の検索装置１０Ａ（または図２の検索装置１０Ｂ）の入力部１１の少なくとも一部の機能と、出力部１４の少なくとも一部の機能を果たす。ユーザは、たとば端末装置３のタッチパネルを操作して検索画面を開き、検索対象として所望の化学物質を入力する。検索対象は、正規表現拡張文字列で特定されてもよいし、標準ＳＭＩＬＥＳ、ＳＭＡＲＴＳ，化学構造式、マルクーシュ構造等の、他の表現形式で特定されてもよい。

【００４４】

化学物質の検索要求は、化学構造の指定とともに、ネットワーク２を介してサーバ装置２００に送信される。

【００４５】

サーバ装置２００の検索処理部２０は、図１の検索装置１０Ａまたは図２の検索装置１０Ｂと同様の機能を果たす。検索処理部２０は、受信した検索要求から化学物質の正規表現拡張文字列を取り出す。検索対象の化学物質が、その他の表現形式で指定されている場合は、正規表現拡張文字列に変換する。正規表現拡張文字列に基づいて、外部の化学物質データベース２６を検索して、該当する化学物質を抽出する。図２のように正規表現拡張文字列をグラフオートマトンに変換する場合は、グラフオートマトンにしたがって、外部の化学物質データベース２６を検索する。グラフオートマトンで定義される遷移状態とマッチングが得られた化学物質が、検索条件に一致する化学物質として抽出される。

【００４６】

データベース２５は、図１のデータベース１５Ａ、または図２のデータベース１５Ｂと同じ機能を果たし、検索の過程で生成された正規表現拡張文字列やグラフオートマトンを、検索により抽出された化学構造式と対応付けて記録する。データベース２５に格納されるデータは、以降の処理で、グラフオートマトンへの変換、マッチング処理等に利用されてもよい。

【００４７】

検索処理部２０による検索結果は、ネットワーク２を介して端末装置３に送信され、端末装置３に表示される。この検索システム１Ｃで用いられるサーバ装置２００も、図３のハードウエア構成で実現可能である。

【００４８】

図５Ａと図５Ｂは、化学構造の正規表現拡張ＳＭＩＬＥＳを例示する図である。比較例として、標準ＳＭＩＬＥＳの文字列を記載する。

【００４９】

例１で、トルエンまたはフェノールを含む物質を検索する場合、ベンゼンの水素原子の一つをメチル基、またはヒドロキシル基で置換したものが検索対象となる。標準ＳＭＩＬＥＳによる検索では、トルエンの文字列と、フェノールの文字列を個別に入力するが、正

規表現拡張ＳＭＩＬＥＳでは、一つの文字列の中で、選択肢の置換基を「｛Ｃ｜Ｏ｝」と表現すればよい。

【００５０】

例２で、ビシクロヘキシルまたはシクロヘキシルシクロペンタンまたはビシクロペンチルを含む物質を検索する場合、標準ＳＭＩＬＥＳによる検索では、シクロヘキシルとシクロペンタンの結合の方向を含めて４通りの文字列を用いる。これに対し、正規表現拡張ＳＭＩＬＥＳでは、シクロヘキシルとシクロペンタンの４通りの組み合わせを表わすのに、２つの集合を含む文字列を用いるだけでよい。

【００５１】

例３で、シクロアルカンの集合を含む物質を検索する場合、標準ＳＭＩＬＥＳでは、メチレン基の数（構造式中のｎは０以上の整数）に応じて、すべての構造を表わす文字列を入力するので、無限範囲の表記は不可能である。これに対し、正規表現拡張ＳＭＩＬＥＳでは、０回以上の繰り返しを表わす記号「＊」を用いて、「Ｃ１Ｃ｛Ｃ｝＊Ｃ１」と表記するだけでよい。ここで、２つの「１」の文字は、分子の同じ位置で連結して炭素の環を形成していることを示すラベルである。

【００５２】

例４で、ベンゼンとアセン類の集合を含む物質を検索する場合、標準ＳＭＩＬＥＳでは直線状に縮合するベンゼン環の数によって、すべての構造を文字列で特定するので、環の数が多くなるほど入力が長くなる。これに対し、正規表現拡張文字列では、０回以上の繰り返しを表わす「＊」を用いて、「ｃ１ｃｃｃ｛ｃ（ｃ１ｃ１）ｃ｝＊ｃｃ１」と表記すればよい。ベンゼン環を含む芳香族の場合、炭素を小文字の「ｃ」で表記している。

【００５３】

上記の例以外にも、直前の集合の１回以上の繰り返しを含む構造を検索したい場合は「＋」を用いて表記すればよいし、所定範囲（ｉ回以上、ｊ回以下）の繰り返しを特定することも可能である。

【００５４】

図６は、正規表現拡張ＳＭＩＬＥＳの文字列からグラフオートマトンへの変換を説明する図である。たとえば、例４のベンゼンとアセン類の集合を含む物資を検索する場合、変換部１２には、正規表現拡張ＳＭＩＬＥＳの文字列「ｃ１ｃｃｃ｛ｃ（ｃ１ｃ１）ｃ｝＊ｃｃ１」が入力される。この正規表現拡張ＳＭＩＬＥＳの文字列は、図６の上段の化学構造式に対応する。

【００５５】

変換部１２は、入力された文字を先頭から順に読み込んで、開始状態からの状態遷移を規定する。入力された正規表現拡張ＳＭＩＬＥＳの文字列の最初の「ｃ」が開始状態ｑ０となる。この最初の「ｃ」は、構造式の繰り返し部分（角括弧の中）を除く頂点に対応する炭素原子である。一例として、左側のベンゼン環の上側の頂点をｑ０とする。

【００５６】

入力文字列の最初の「ｃ」は直後に数字の「１」を伴い、開始点で環が閉じられることが示されている。入力文字列の「ｃ１」は、グラフオートマトンの状態「ｑ０（ｃ（：，@））」に変換される。「ｃ」は芳香族炭素を表わし、コロン「：」は、１つの芳香族炭素が結合していることを表わす。「@」はｑ０への戻りパスがあることを示している。

【００５７】

この例では、説明を簡単にするために結合の種類（単結合、多重結合）を特定していないが、コロン「：」に替えて、単結合を表わす記号「－」や、二重結合を表わす記号「＝」を用いて結合の種類を表わしてもよい。

【００５８】

入力文字列で、「ｃ１」に続いて「ｃ」が記述されている。開始状態ｑ０には、次の頂点「ｃ」への遷移パスと、反対方向からのｑ０への戻りパスがある。ｑ０からｑ１への遷移と、ｑ０への戻りパスの存在は、グラフオートマトンで

ｑ０（ｃ（：，@））－＞ｃ（ｑ１（：），ｐ１（@））

と規定される。

【００５９】

入力文字列が４番目の「ｃ」まで読み込まれると、状態は、ｑ１、ｑ２、ｑ３と順に遷移する。ここまでが、グラフオートマトンの１行目から３行目に規定される遷移である。

【００６０】

入力文字列の４番目の「ｃ」の後に、０回以上繰り返される集合「｛ｃ（ｃ１ｃ１）ｃ｝＊」が記述されている。したがって、状態ｑ３には、繰り返し回数０のときの遷移（集合の後ろの「ｃ」へのパス）と、繰り返しがあるときの遷移（集合内の「ｃ」へのパス）の２通りの遷移が存在する。グラフオートマトンでは、この２通りの遷移が以下のように規定される。

【００６１】

ｑ３（ｃ（：））－＞ｃ（ｑ４（：））
ｑ３（ｃ（：））－＞ｃ（ｑ８（：））

状態ｑ３から状態ｑ４への遷移は、構造式の角括弧の中の繰り返しへの遷移である。状態ｑ３から状態ｑ８への遷移は、角括弧の外の頂点への遷移である。

【００６２】

状態ｑ４では、分岐が行われる。集合｛ｃ（ｃ１ｃ１）ｃ｝の中で、分岐を表わす記号「（）」が記述されており、括弧内の最初の「ｃ１」への分岐と、括弧の後ろの「ｃ」への分岐である。この分岐は、状態ｑ４から状態ｑ５への遷移パスと、状態ｑ４から状態ｑ７への遷移パスとして、グラフオートマトンで、

ｑ４（ｃ（：，：））－＞ｃ（ｑ５（：），ｑ７（：））

と規定される。

【００６３】

括弧内の最初の「ｃ１」の後に、もう一つ「ｃ１」が続く。すなわち、状態ｑ５には、入力文字列の最初の「ｃ１」に対応する状態ｑ０に戻るパスｐ１（：）と、次の「ｃ１」に対応する状態ｑ６への遷移パスがある。

【００６４】

これは、グラフオートマトンで

ｑ５（ｃ（：，：））－＞ｃ（ｐ１（：），ｐ６（：））

と規定される。状態ｑ６では、括弧内の２つめの芳香族炭素「ｃ１」への戻りパスを待ち受ける。

【００６５】

一方、状態ｑ７で、繰り返しが続く場合は、状態ｑ４（集合内の最初の「ｃ」）に戻る。繰り返しが有限回数で指定されている場合は、繰り返しの終了により状態ｑ８（集合の後ろの最初の「ｃ」）に遷移する。これらの遷移は、グラフオートマトンの９行目と１０行目に規定される。

【００６６】

入力文字列の最後の「ｃ１」は、状態ｑ９として定義され、状態ｑ６への戻りパスｐ１をたどる（グラフオートマトンの１１行目と１２行目）。

【００６７】

別の例として、図５Ｂの例３のシクロアルカンの集合を検索する場合は、変換部１２は入力文字列「Ｃ１Ｃ｛Ｃ｝＊Ｃ１」を先頭から読み込んで、以下のようなグラフオートマトンを生成する。

【００６８】

ｑ０（Ｃ（－，@））－＞Ｃ（ｑ１（－），ｐ１（@））
ｑ１（Ｃ（－））－＞Ｃ（ｑ２（－））
ｑ１（Ｃ（－））－＞Ｃ（ｑ３（－））
ｑ２（Ｃ（－））－＞Ｃ（ｑ３（－））
ｑ２（Ｃ（－））－＞Ｃ（ｑ２（－））
ｑ３（Ｃ（－））－＞Ｃ（ｐ１（－））

このように、グラフオートマトンへの変換は、変換部１２に入力された正規表現拡張文字列を先頭から順に読み込み、グラフに見立てた化学構造式の頂点間の遷移状態を規定する処理である。

【００６９】

図７は、検索処理部１３による検索結果の一例を示す図である。この例では、検索対象として「ベンゼンとアセン類の集合」が指定されており、１つ以上のベンゼン環が直線状に縮合した構造を含むすべての化学物質が抽出される。検索番号２９３では、３つのベンゼン環が直線状に縮合したアセトラセンを含む物質がリストされ、検索番号３００では、２つのベンゼン環が直線状に縮合したナフタレンを含む物質がリストされている。

【００７０】

マッチング処理にグラフオートマトンを用いる場合、グラフオートマトンで規定された状態遷移を満たす化学物質が特定される。任意のデータベースに格納されている既存の化学物質のノードを順にたどって、グラフオートマトンで規定された状態遷移が満たされるどうかを判断するので、化学物質に含まれる一部分が、指定された化学物質と一致する場合も、正確に抽出することができる。

【００７１】

検索処理部１３による検索結果は、出力部１４によって出力され、たとえば表示装置の表示画面に表示される。図７の出力例で、灰色の枠内の番号は検索結果の通し番号、白枠の番号は、使用された化学物質データベースでの登録番号である。

【００７２】

図７では、「ベンゼンとアセン類の集合」に含まれる物質を有するすべての化学物質が抽出され、表示されている。しかし、最小一致の物質や、最大一致の物質を検索したい場合がある。たとえば、ベンゼン環を１つでも含む物質をすべて検索したい場合や、最大数のベンゼン環を含むアセン類を検索したい場合などである。

【００７３】

図８は、入力部１１のインターフェースである検索画面１１１の一例を示す。検索対象入力ボックス１１２の他に、検索条件選択ボックス１１３が表示されている。ユーザは、検索対象入力ボックス１１２に、「ベンゼンとアセン類の集合」を表わす正規表現拡張ＳＭＩＬＥＳ「c1ccc{c(c1c1)c}*cc1」を入力し、検索条件として「最小一致」を選択する。入力部１１への検索対象の入力は、必ずしも正規表現拡張ＳＭＩＬＥＳでなくてもよいが、正規表現拡張ＳＭＩＬＥＳで入力する場合は、入力パターンが短く、入力作業が簡単になる。

【００７４】

この入力文字列は、変換部１２によってグラフオートマトンに変換されてもよい。検索処理部１３は、グラフオートマンにしたがって、データベース中の化学物質をひとつずつ調べる。検索条件として「最小一致」が選択されているので、化学物質中に１つでもベンゼンが含まれていれば、検索条件に合致すると判定される。たとえば、図６のオートマトンで「ｑ５（ｃ（：，：））－＞Ｃ（ｐ１（：）、ｑ６（：））まで遷移できたところで、「一致」と判断して、次の化学物質の検索に進んでもよい。

【００７５】

検索条件は、図８の例に限定されず、「完全一致」、「類似構造検索」などの検索条件を選択可能にしてもよい。「類似構造検索」とは、たとえば、入力文字列で記述される化学構造と、構成元素や置換基の種類、位置等が異なっていても構造が類似する化学物質の検索である。化学物質において、グラフオートマトンで規定される状態と元素の種類が異なっていても同じ遷移をたどる場合は、類似物資として抽出される。

【００７６】

図９は、「最小一致」の検索条件が選択されたときの、「ベンゼンまたはアセン類の集合」を含む化学物質の検索結果の表示例である。「最小一致」が検索されたとき、たとえば直線状に縮合する複数のベンゼン環のうち、１つだけを実線で表示し、他のベンゼン環を破線で表示してもよい。あるいは、１つのベンゼン環の色を変える等、任意のハイライ

ト表示が可能である。

【００７７】

図１０は、グラフオートマトンの別の変換例を示す図である。ここでは「トルエンまたはフェノール」を含む物質（図５Ａの例１）を検索する。検索装置１０の変換部１２は、入力された正規表現拡張ＳＭＩＬＥＳの文字列「｛Ｃ｜Ｏ｝ｃ１ｃｃｃｃｃ１」を先頭から順に読み込む。文字列の先頭に、置換基の選択肢を表わす集合｛Ｃ｜Ｏ｝が記述されているので、グラフオートマトンで開始状態ｑ０と開始状態ｑ１が生成される。開始状態ｑ０は、炭素原子に対応するノードを表わしている。もうひとつの開始状態ｑ１は、酸素原子に対応するノードを表わしている。

【００７８】

集合を表す文字列「｛Ｃ｜Ｏ｝」の直後に、「ｃ」が記述されている。グラフオートマトンで、開始状態ｑ０から状態ｑ２への遷移と、開始状態ｑ１から状態ｑ２への遷移が規定される。状態ｑ２は、芳香族炭素に対応するノードである。

【００７９】

入力文字列において、集合の直後の「ｃ」は、後ろに「１」を伴うので、この「ｃ」でベンゼン環が閉じられ、逆方向からのパスがあることが示されている。文字列では、「１」に続いて「ｃ」が記述されている。すなわち、状態ｑ２は、状態ｑ３に遷移可能であるとともに、逆方向からのパスｐ１が待ち受け可能（「@」）である。グラフオートマトンで、

ｑ２（ｃ（：，@））－＞ｃ（ｑ３（：），ｐ１（@））

の遷移が規定される。ここでは、図６のグラフオートマトンと整合をとるために、芳香族炭素ｃの結合を「：」で示しているが、単結合か二重結合かに応じて「－」と「＝」を使い分けてもよい。

【００８０】

以下、状態ｑ３から状態ｑ７へと順番に遷移し、入力文字列の最後の「ｃ１」で、状態ｑ７から戻りパスｐ１でｑ２に戻る。

【００８１】

このようにして生成されたグラフオートマトンは、正規表現拡張文字列及び構造式と対応付けて、データベース１５に保存されてもよい。

【００８２】

検索処理部１３は、データベース１５及び／または外部の化学物質データベースを参照して、グラフオートマトンに従って化学物質をひとつずつ調べる。グラフオートマトンで定義される状態遷移を満たす化学物質が抽出され、出力される。

【００８３】

図１１は、データベース１５Ｂ（または２５）に保存されるグラフオートマトンの記録例を示す。生成されたグラフオートマトンを、正規表現拡張文字列と、化学構造式とに関連付けて保存する。これ以外にも、「ベンゼン」、「シクロアルカン」などという名称や、分子式等を関連付けて保存してもよい。

【００８４】

変換部１２は、正規表現拡張文字列が入力されたときに、データベース１５Ｂを参照して、すでに対応するグラフオートマトンが保存されている場合は、保存されたグラフオートマトンを読み出して検索処理部１３に渡せばよい。これによって検索時間を短縮することができる。

【００８５】

図１２は、化学構造の検索方法のフローチャートである。この処理フローは、検索システム１Ａの検索装置１０Ａ、または検索システム１Ｂの検索装置１０Ｂで実行されてもよいし、検索システム１Ｃのサーバ装置２００によって実行されてもよい。

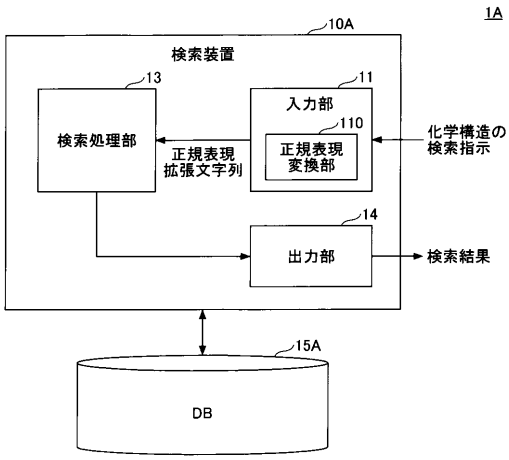
【００８６】

まず、化学構造の正規表現拡張文字列を取得する（Ｓ１１）。検索装置１０Ａ，１０Ｂ、またはサーバ装置２００に、直接、正規表現拡張文字列が入力されてもよいし、その他

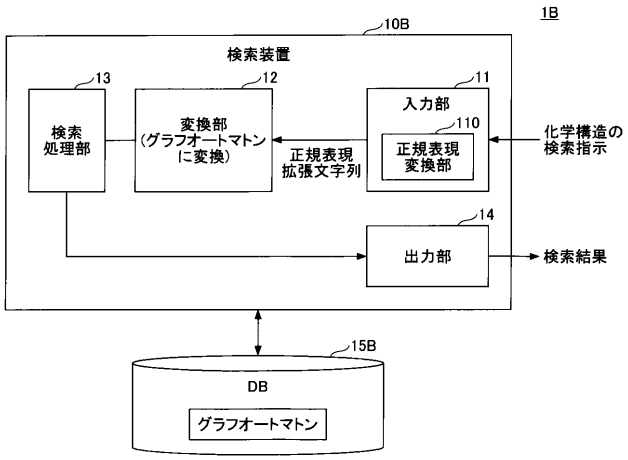
の表現形式で化学構造が特定されている場合は、正規表現拡張文字列に変換する。
【００８７】
図１の検索装置１０Ａを用いる場合は、ステップ１３に飛んで検索処理を行う。図２の検索装置１０Ｂを用いる場合は、入力された正規表現拡張文字列を、グラフオートマトンに変換する（Ｓ１２）。グラフオートマトンへの変換処理は、上述した通り、入力された文字列を最初から順に読み込み、グラフに見たてた化学構造のノードからノードへの遷移を規定する。
【００８８】
入力された正規表現拡張文字列、または生成されたグラフオートマトンにしたがって、検索処理を行う（Ｓ１３）。検索処理では、任意の化学物質データベースに記録されている化学物質について、正規表現拡張文字列に一致するか否か、またはグラフオートマンで規定される遷移を満たすか否かが判定される。グラフオートマトンで規定される遷移を満たす物質は、検索条件に一致すると判断される。
【００８９】
最後に、検索結果を出力する（Ｓ１４）。図１の検索システム１Ａまたは図２の検索システム１Ｂの場合は、検索装置１０Ａまたは１０Ｂの表示装置等の出力部１４に検索結果を表示する。図４の検索システム１Ｃの場合は、サーバ装置２００から端末装置３に、検索結果を送信し、端末装置３の表示画面に検索結果が表示される。
【００９０】
化学構造を正規表現拡張文字列で表現することで、置換、集合、繰り返し等を含む多様な構造を一つの文字列で簡単に表現することができる。
【００９１】
正規表現拡張文字列をグラフオートマトンに変換する場合は、既存のデータベースに保存されている化学物質とのマッチング判定が容易になり、検索速度が速くなる。
【００９２】
検索の過程で、生成されたグラフオートマトンを入力文字列と化学式に関連付けて保存することで、グラフオートマトンのデータベースを構築することができる。グラフオートマトンのデータベースを、化学物質の検索に利用することもできる。
【００９３】
実施形態の検索をプログラムで実現する場合は、プログラムに記述された以下の手順をコンピュータによって実行する。
（ａ）化学構造を表わす分子記述言語に正規表現を適用して拡張した正規表現拡張文字列を取得する手順；及び
（ｂ）正規表現拡張文字列をもとに、データベースから該当する化学物質を抽出する手順。
【００９４】
グラフオートマトンを利用する場合は、上記の手順に加えて、正規表現拡張文字列をグラフオートマトンに変換する手順をコンピュータに実行させてもよい。この場合は、グラフオートマトンにしたがってデータベースから該当する化学物質を抽出する。
【００９５】
これによって、検索装置１０Ａ，検索装置１０Ｂ、またはサーバ装置２００を実現することができる。
【符号の説明】
【００９６】
１Ａ、１Ｂ、１Ｃ　検索システム
２　ネットワーク
３Ａ～３Ｎ　端末装置
１０Ａ，１０Ｂ　検索装置
１１　入力部
１２　変換部

１３　検索処理部
１４　出力部
１５Ａ、１５Ｂ，２５　データベース（記憶部）
２０　検索処理部
１１１　検索画面
１１２　検索対象入力ボックス
１１３　検索条件選択ボックス
２００　サーバ装置

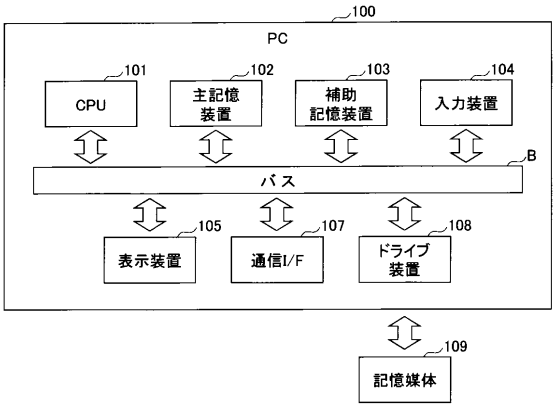
【図１】



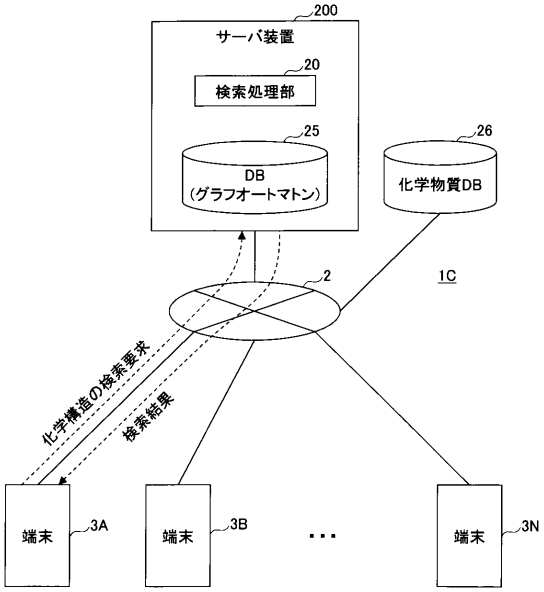
【図２】



【図３】



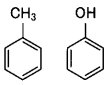
【図４】



【図５Ａ】

化学構造の正規表現拡張文字列の例

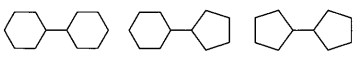
1. トルエンまたはフェノール



実施形態の入力文字列
:{C|O}c1ccccc1

c.f. 既存の入力文字列
:{Cc1ccccc1,Oc1ccccc1}

2. ビシクロヘキシルまたはシクロヘキシルシクロペンタンまたはビシクロペンチル



実施形態の入力文字列
:{C1CCCC1|C1CCC1}{C1CCCC1|C1CCC1}

c.f. 既存の入力文字列
:{C1CCCC1C1CCCC1,C1CCCC1C1CCC1,
C1CCC1C1CCCC1,C1CCC1C1CCC1}

【図５Ｂ】

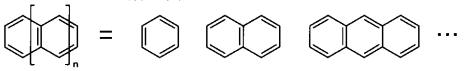
化学構造の正規表現拡張文字列の例

3. シクロアルカンの集合

実施形態の入力文字列
:C1C{C}*C1

c.f. 既存の入力文字列
:{C1CC1,C1CCC1,C1CCCC1,C1CCCCC1,・・・}

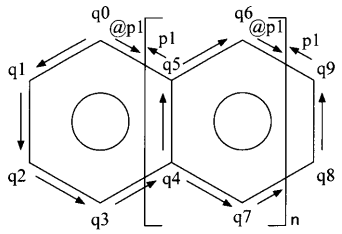
4. ベンゼンとアセン類の集合



実施形態の入力文字列
: c1ccc{c(c1c1)c}*cc1

c.f. 既存の入力文字列
: {c1ccccc1,c1cccc(c1c1)ccc1,c1cccc(c1c1)cc(c1c1)ccc1,・・・}

【図６】

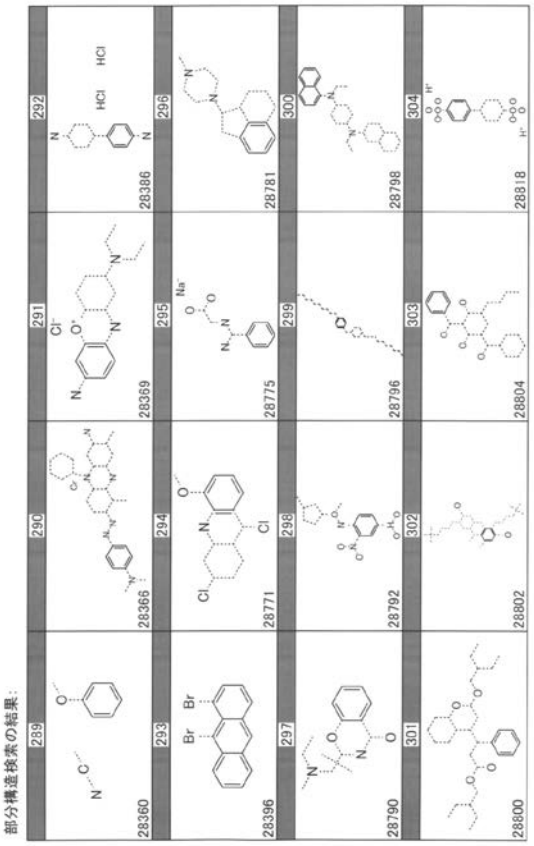


入力された正規表現拡張文字列： c1ccc{c(c1c1)c}*cc1

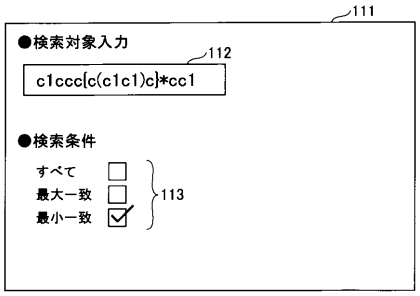
⇩ グラフオートマトンへの自動変換

q0(c(:,@)) -> c(q1(:),p1(@))
q1(c(:)) -> c(q2(:))
q2(c(:)) -> c(q3(:))
q3(c(:)) -> c(q4(:))
q3(c(:)) -> c(q8(:))
q4(c(:,:)) -> c(q5(:),q7(:))
q5(c(:,:)) -> c(p1(:),q6(:))
q6(c(@)) -> c(p1(@))
q7(c(:)) -> c(q4(:))
q7(c(:)) -> c(q8(:))
q8(c(:)) -> c(q9(:))
q9(c(:)) -> c(p1(:))

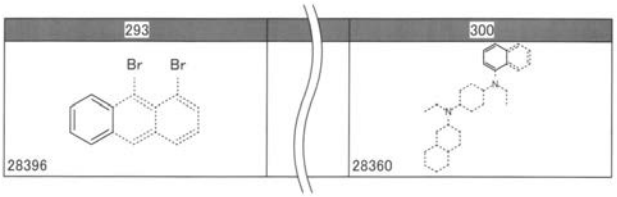
【図７】



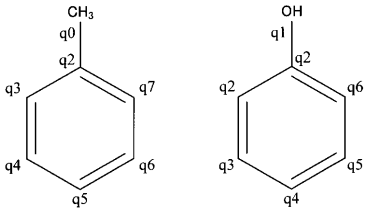
【図８】



【図９】



【図１０】

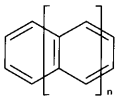


入力文字列： {C|O}c1ccccc1

⇩ グラフオートマトンへの自動変換

q0(C(-)) -> C(q2(-))
q1(O(-)) -> O(q2(-))
q2(c(:,@)) -> c(q3(:),p1(@))
q3(c(:)) -> c(q4(:))
q4(c(:)) -> c(q5(:))
q5(c(:)) -> c(q6(:))
q6(c(:)) -> c(q7(:))
q7(c(:)) -> c(p1(:))

【図１１】

正規表現拡張文字列	化学構造式	グラフオートマトン
c1ccccc1		q0(c(:,@)) → c(q1(:), p1(@)) q1(c(:)) → c(q2(:)) q2(c(:)) → c(q3(:)) q3(c(:)) → c(q4(:)) q4(c(:)) → c(q5(:)) q5(c(:)) → c(q6(:))
C1C{C}*C1		q0(C(-,@)) → C(q1(-),p1(@)) q1(C(-)) → C(q2(-)) q1(C(-)) → C(q3(-)) q2(C(-)) → C(q3(-)) q2(C(-)) → C(q2(-)) q3(C(-)) → C(p1(-))
c1ccc{c(c1c1)c}*cc1		q0(c(:,@)) → c(q1(:),p1(@)) q1(c(:)) → c(q2(:)) q2(c(:)) → c(q3(:)) q3(c(:)) → c(q4(:)) q3(c(:)) → c(q8(:)) q4(c(:,:)) → c(q5(:),q7(:)) q5(c(:,:)) → c(p1(:),q6(:)) q6(c(@)) → c(p1(@)) q7(c(:)) → c(q4(:)) q7(c(:)) → c(q8(:)) q8(c(:)) → c(q9(:)) q9(c(:)) → c(p1(:))

【図１２】