

形式文法を用いて化学構造式中の示性式を 自動認識する化学構造式 OCR の提案

佐藤 文恭^{1,a)} 藤芳 明生^{2,b)}

受付日 2015年11月26日, 採録日 2016年8月10日

概要: 本稿では, 形式文法を用いて化学構造式中の示性式を自動認識する方法を提案する. 示性式とは, 文字のみで構成される化学構造式であり, 一般的な化学構造式の一部として現れることがある. そのため, 化学構造式 OCR を開発するためには, 示性式における原子間の結合情報を得る必要がある. 従来の OCR アプリケーションでは, 事前に示性式の構造を登録した辞書を用いることで, 化学構造式画像の認識を行っている. しかし, この方法では辞書に登録されていない示性式は処理することができない. そこで, 形式文法を用いて示性式の構文と意味を定義し, 示性式の認識を文法の構文解析処理として行うことを提案する.

キーワード: 化学構造式 OCR, 示性式, 形式文法, 構文解析

Proposal of the Uses of a Formal Grammar to Recognize Condensed Structural Formulas for Optical Chemical Structure Recognition

FUMIYASU SATO^{1,a)} AKIO FUJIYOSHI^{2,b)}

Received: November 26, 2015, Accepted: August 10, 2016

Abstract: In this paper, we propose a method to recognize condensed structural formulas in chemical structural formulas automatically using a formal grammar. A condensed structural formula is a chemical structural formula that consists of only characters. It appears as a part of general chemical structural formula. Therefore, in order to develop optical chemical structure recognition, we need to obtain the information of bonds between atoms in condensed structural formulas. Conventional optical chemical structure recognition applications use a predefined dictionary which stores the structure of condensed structural formulas. However, condensed structural formulas not stored in the dictionary cannot be processed. So, we define syntax and semantics of condensed structural formulas using a formal grammar and, thus, a recognition of condensed structural formulas can be done as a parsing process of the grammar.

Keywords: optical chemical structure recognition, condensed structural formulas, formal grammar, parsing

1. はじめに

化学や薬学の分野において, 日々膨大な数の化合物が発見され, それらの化合物の化学構造式が論文や特許出願書

類として公開されている. 化学構造式の例を図 1 に示す. 化学構造式とは, 原子を表す記号と原子間の結合を意味する線を用いて化合物の分子構造を示すものである. 論文や特許出願書類中の化学構造式画像の分子構造を正確に自動認識するために, 化学構造式 OCR の開発が行われている.

化学構造式 OCR の開発において, 化学構造式の一部として現れる示性式の処理を考慮する必要がある. 示性式とは, 文字のみで構成される化学構造式である. たとえば, 図 1 中では, “N”, “OH”, “COOH”, “SO₃Na” などが示性式である. 示性式は原子間の結合が線によって明示されてい

¹ 茨城大学大学院理工学研究科
Graduate School of Science and Engineering, Ibaraki University, Hitachi, Ibaraki 316-0033, Japan

² 茨城大学工学部
Faculty of Engineering, Ibaraki University, Hitachi, Ibaraki 316-0033, Japan

a) 14nm710n@vc.ibaraki.ac.jp

b) akio.fujiyoshi.cs@vc.ibaraki.ac.jp

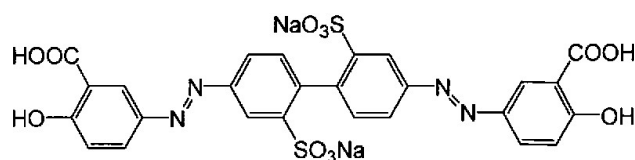


図 1 化学構造式

Fig. 1 Chemical structural formula.

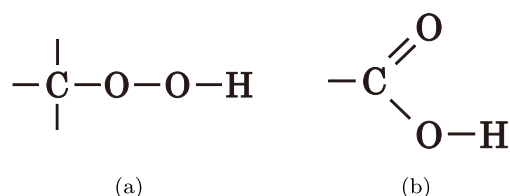


図 2 カルボキシ基の分子構造の認識

Fig. 2 Recognition of molecular structure of carboxy group.

ないにもかかわらず，示性式の分子構造は一意に認識することができるように書かれている．たとえば，カルボキシ基 “COOH” の分子構造について考えてみる．炭素 “C” の手の数は 4 本，“O” の手の数は 2 本，“H” の手の数は 1 本であることが分かっている．図 2(a) のように，炭素と 1 つ目の酸素が単結合し，1 つ目の酸素と 2 つ目の酸素が単結合し，2 つ目の酸素と水素が単結合しているとなると，化学構造式の主構造と結合している 1 本を除いても炭素の手の数が 2 本余ってしまう．一方，図 2(b) のように，炭素と 1 つ目の酸素が二重結合し，炭素と 2 つ目の酸素が単結合し，2 つ目の酸素と水素が単結合しているとなると，化学構造式の主構造と結合している 1 本を除くと手の数は余らない．よって，カルボキシ基の分子構造は図 2(a) のように認識するよりも図 2(b) のように認識する方がふさわしいと分かる．このように，示性式の分子構造を正しく認識するためには化学に関するある程度の知識が必要である．

既存の化学構造式 OCR アプリケーションとして Mol-Rec [1], [2], [3], [4], OSRA [5], ChemInfty [6], [7], [8] などが存在する．それらは化学構造式中の示性式の構造を得るために，事前に示性式の構造を登録した辞書を用いて化学構造式画像の認識を行っている．しかし，この方法では辞書に登録されていない示性式は処理することができない．したがって，より一般的な方法が求められている．

本稿では，形式文法を用いて示性式の構文と意味を定義し，その文法で示性式を認識する方法を提案する．提案手法は 2 つの段階で構成される．1 つ目の段階では，省略された結合を表す文字の追加および分岐や連結を示す表記の拡張を行い，「展開文字列」を作成する．2 つ目の段階では，展開文字列に文法を用いて構文解析し，分子構造の情報を得る．示性式の文法的複雑さと構文解析の容易さを考慮し，展開文字列の構文は，文脈自由文法を用いて記述した．

提案手法を評価するために，JAPIC の日本の医薬品構造

表 1 化学構造式中で使用される略称

Table 1 Abbreviations used in chemical structural formulas.

略称	基の名前	構造式
Me	メチル基	CH ₃
Et	エチル基	CH ₃ CH ₂
Pr	プロピル基	CH ₃ CH ₂ CH ₂
Bu	ブチル基	CH ₃ CH ₂ CH ₂ CH ₂
N ₃	アジ基	NN ⁺ N ⁻
Ph	フェニル基	C ₆ H ₅
Bn	ベンジル基	CH ₂ C ₆ H ₅

式集 [9] の中から高分子を取り除いた 1,427 個の化学構造式を用いてテストを行った．その結果，94% (1,346/1,427) の化学構造式に登場する示性式を正確に認識できたことを確認した．

2. 化学構造式

化学構造式とは，化合物の分子を構成している原子が互いにどのように結合しているのかを，原子を表す記号と結合を表す線を用いて二次元平面上に模式的に示した式である．“-”，“=”，“≡” はそれぞれ単結合，二重結合，三重結合を表す結合記号である．化学構造式では，その分子の特徴を表す基や構造部分を強調するためにそれらを略称を用いて表記することもある．頻繁に使用される略称を表 1 に示す．化学構造式中の骨格となる炭化水素部分の炭素原子 “C” と水素原子 “H” は省略した形で表記することが多い．その場合は炭素原子間を結ぶ線の端点および屈折点，交差点が炭素原子を表す．また，省略された炭素原子に結合している水素原子のほとんどは結合を含めて省略される．

2.1 原子価

原子価とは，ある原子が他の何個の原子と結合することができるかを表した数である．以下，原子価を「手の数」と呼ぶことにする．手の数は，水素原子は 1 本，酸素原子は 2 本，炭素原子は 4 本というように，ほとんどの原子は決まった手の数を持つ．しかしながら，原子価が 2 価，4 価，6 価の場合がある硫黄のように，何通りかの手の数をとりうる原子も存在する．

2.2 示性式

示性式とは，原子を表す記号と括弧および結合の種類を表す記号，電荷を表す記号，量を表す数字から構成され，それぞれの原子の原子価を考慮することで一意に分子構造を認識することができるように書かれた化学構造式である．たとえば，酢酸の分子式 “C₂H₄O₂” は，構造が明確ではないのに対し，酢酸の示性式 “CH₃COOH” は，メチル基 “-CH₃” とカルボキシ基 “-COOH” からなる構造であることが明確である．示性式を用いることにより化学構造式の描画領域を節約できるため，示性式は一般的な化学構造式の一部とし

て現れることがある。

示性式は以下のように再帰的に定義することができる.

- 原子を表す文字列は、示性式である。(例：“C”，“H”，“Cl”，“Na”)
- 官能基を表す文字列は、示性式である。(例：“Me”，“Ph”)
- 下付き数字を持つ原子を表す文字列は、示性式である。(例：“H₂”，“Cl₄”)
- 上付きの(数字をとともなうこともある)+または-を持つ原子を表す文字列は、示性式である。(例：“H⁺”，“O⁻”，“Ca²⁺”)
- 括弧で括られた示性式は、示性式である。(例：“(OH)”，“(CH₃)”)
- 下付き数字を持つ括弧で括られた示性式は、示性式である。(例：“(OH)₂”，“(CH₂)₁₀”)
- 示性式 x , y に対し, xy , $x-y$, $x=y$, $x\equiv y$ は、示性式である。(例：“CH₃CH₃”，“CH₃-CH₃”，“CH₃=CH₃”，“CH₃≡CH₃”)

図1の化学構造式中には、“N”，“OH”，“HO”，“COOH”，“HOOC”，“SO₃Na”，“NaO₃S”の示性式が存在する．示性式中では一般に，水素原子“H”は省略されない．原子間の結合は，通常は省略されるが，必要に応じて化学構造式と同様の結合記号が使用される．括弧は，構造の分岐または繰返しを表すために使用される．原子や括弧の後の数字は，繰返しの回数を表す．原子の上付きの“+”または“-”は，電荷を表す．

示性式の分子構造を認識するためには化学に関するある程度の知識が必要となる．特に以下のような点を判断しなくてはならない．

- (1) 示性式中の括弧は構造の分岐だけではなく、構造の繰返しを表すためにも使用されるため、括弧がどちらの意味で使用されているのか判断する必要がある。
- (2) 複数の原子価を持つ原子に対し、どの原子価を適用するのが最もふさわしいのか判断する必要がある。
- (3) 化学構造式の主構造への接続方向によって表記の方向が変わることがあるため、表記の方向を判断する必要がある。

(1) の判断が必要となる例を図 3 に示す。括弧が構造の分岐または繰返ししのどちらの意味で使用されているのか化学の知識なく判断することは不可能である。通常、括弧内の原子の手の数を調べることで判断することができる。まず、図 3 (a) 中の括弧内の“CH₃”の手の数を計算する。炭素原子“C”は手の数が 4 本、水素原子“H”は手の数が 1 本

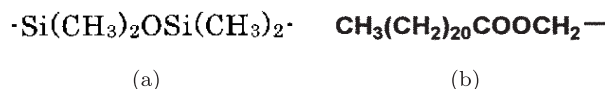


図 3 (a) 構造の分岐と (b) 構造の繰返し

Fig. 3 (a) Branch of structure and (b) repeat of structure.

である。炭素原子に3つの水素原子がそれぞれ単結合していることから、括弧内では炭素原子の手の数が1本余っている状態であることが分かる。余った1本はそれより前に記述されている原子に単結合していると考えられるため、この括弧が表すのは構造の分岐であると推測できる。

次に、図 3(b) 中の括弧内の“CH₂”の手の数も同様に計算する。炭素原子に 2 つの水素原子がそれぞれ単結合していることから、括弧内では炭素原子の手の数が 2 本余っている状態であることが分かる。一般的に、括弧内と括弧外の原子が二重結合、または三重結合をする場合は、それぞれを表す結合記号“=”, “≡”を用いて明示される。図 3(b) 中の括弧内に結合記号は記されていないため、余った 2 本のうちの 1 本がそれより前に記述されている原子に単結合し、残りの 1 本は構造の繰返しとして次の原子に単結合していると考えられる。よって、この括弧が表すのは構造の繰返しであると推測できる。

(2) の判断が必要となる例を図 4 に示す。図 4 は、2 価、4 価、6 価の原子価を持つ硫黄原子と 1 価の原子価を持つフッ素原子からなる 3 種類の化学構造式である。示性式 SF_2 の場合は 2 つのフッ素原子と単結合していると考えられるため、硫黄原子の手の数は 2 本であることが分かる。同様に、示性式 SF_4 の場合は硫黄原子の手の数は 4 本、示性式 SF_6 の場合は硫黄原子の手の数は 6 本であることが分かる。このように、示性式中の複数の手の数をとりうる原子の手の数は、その原子に結合していると考えられる周囲の原子の手の数を計算することで判断することが可能である。

しかし例外的に、手の数の計算だけでは硫黄原子の手の数を判定することが困難な場合も存在する．図 5 に硫黄原子の手の数が 2 本，4 本，6 本のすべての場合で分子構造内の手の数の計算が成り立つ例として，“ SO_3H ” の分子構造を示す．図 5(a) では，硫黄原子に対して 3 つの酸素原子

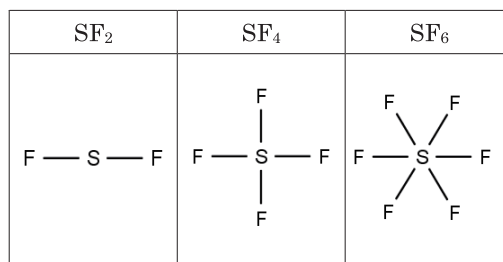


図 4 2 価, 4 価, 6 価の原子価を持つ硫黄原子

Fig. 4 Sulfur has a valence of 2, 4 or 6.

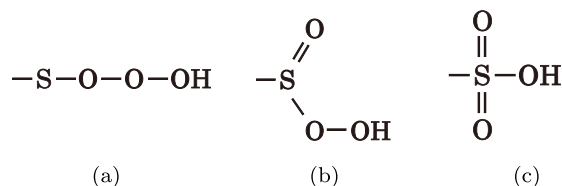


図 5 “SO₃H” の分子構造

Fig. 5 Molecular structure of "SO₃H".

の1つ目だけが単結合しており、硫黄原子の手の数は化学構造式の主構造と結合している1本を含めて2本である。図5(b)では、硫黄原子に対して3つの酸素原子の1つ目が二重結合し、2つ目が単結合しており、硫黄原子の手の数は化学構造式の主構造と結合している1本を含めて4本である。図5(c)では、硫黄原子に対して3つの酸素原子の1つ目と2つ目がそれぞれ二重結合し、3つ目が単結合しており、硫黄原子の手の数は化学構造式の主構造と結合している1本を含めて6本である。以上の3つの分子構造が考えられるが、一般的に“SO₃H”はスルホ基と呼ばれる官能基であり、図5(c)の分子構造が正しいと理解される。示性式の認識にはこのような知識も必要となることもある。

(3)の判断が必要となる例を図6に示す。図6(a)は化学構造式の主構造の右側に描かれた示性式の場合である。この場合、一般的な示性式と同様に左から右に読み取ることで、分子構造を正しく認識することができる。図6(b)と図6(c)は化学構造式の主構造の左側に描かれた示性式である。この場合、図6(b)のように一般的な示性式とは逆に右から左に読み取ることで分子構造を正しく認識することができることが多いが、稀に図6(c)のように括弧の外は逆の並びになっているにもかかわらず、括弧の中は左から右の順に並んでいることもある。したがって、主構造の左側に描かれた示性式の分子構造を認識する際には上記の3つのパターンを考慮する必要がある。

2.3 完全構造式

化合物中のすべての原子間における結合を表す線を省略せずに明記した化学構造式を、完全構造式と呼ぶことにする。完全構造式の例として、図1の化学構造式を完全構造式にしたものを図7に示す。示性式は、一般に原子間の結合を表す線が省略されているため、示性式を持つ化学構造式は完全構造式ではない。化学構造式の分子構造を認識するということは、示性式などの原子間の結合情報が明確では

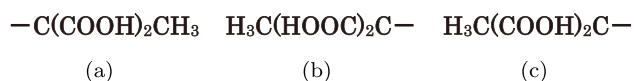


図6 表記の方向が異なる例

Fig. 6 Example of difference of directions of notation.

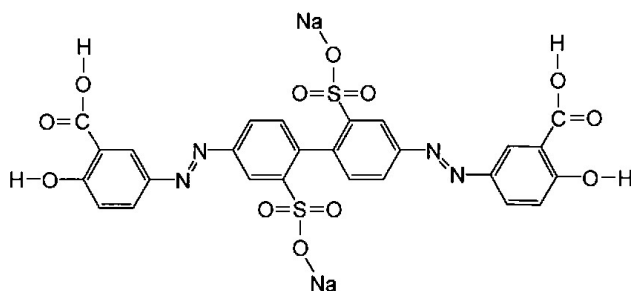


図7 完全構造式

Fig. 7 Complete structural formula.

ない部分を解析し、完全構造式を得ることであるといえる。

3. 形式文法

形式文法とは、文字列の集合を定義するために生成規則を用いて文字列を生成する仕組みである。形式文法 G は、以下のように4つ組として定義される [10]。

$G = (V_N, V_T, P, S)$, ここで、

- V_N は空でない非終端記号の有限集合である。
- V_T は空でない終端記号の有限集合である。
- $S \in V_N$ は開始記号である。
- P は $\alpha \rightarrow \beta$ の形の生成規則の有限集合である。ここで、 α は長さ1以上の終端記号と非終端記号からなる文字列であり、 β は長さ0以上の終端記号と非終端記号からなる文字列である。

形式文法 G が生成する言語とは、開始記号 S から始め、生成規則を適用し文字列を書き換えることを繰り返し、最終的に得られる終端記号からなる文字列の集合である。

3.1 文脈自由文法

文脈自由文法とは、 P に含まれるすべての生成規則の左辺が1つの非終端記号であるような形式文法である。文脈自由文法の例として、言語 $\{a^n b^n | n \geq 0\}$ を生成する文脈自由文法 G を以下に示す。

$G = (N, \Sigma, P, S)$

- $N = \{S\}$
- $\Sigma = \{a, b\}$
- $P = \{S \rightarrow ab, S \rightarrow aSb\}$

文字列“aaabbb”は G によって以下のように生成される。

$S \Rightarrow aSb \Rightarrow aaSbb \Rightarrow aaabbb$

3.2 構文木

構文木とは、文脈自由文法が文字列を生成する様子を木構造で表現したものである。前節の文字列“aaabbb”を文脈自由文法 G が生成するときの構文木を、図8に示す。構文木を根から葉に向かってたどることで、どのように生成規則が適用されて文字列が生成されているのか知ることができる。与えられた文字列から構文木を作る作業を構文解析と呼ぶ。

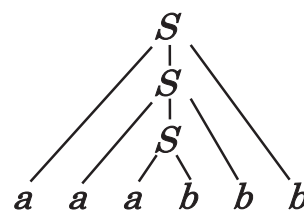


図8 構文木

Fig. 8 Parse tree.

4. 提案手法

本稿では形式文法を用いて示性式の構文と意味を定義し、その文法で示性式を認識する方法を提案する。提案手法は2つの段階で構成される。1つ目の段階では、省略された結合を表す文字の追加および分岐や連結を示す表記の拡張を行い、「展開文字列」を作成する。2つ目の段階では、展開文字列に文法を用いて構文解析し、構文木を生成する。この構文木から、完全構造型を得ることができる。

4.1 展開文字列の作成

1つ目の段階では、示性式を展開文字列に変換する。例として、化学構造式中の示性式 “-CMe₂N₃” を展開文字列に変換する様子を以下に示す。

$$\begin{aligned}
 & *CMe_{-}\{2\}N_{-}\{3\} & (1) \\
 \Rightarrow & *C(CH_{-}\{3\})_{-}\{2\}(NN^{+}\{+ \}N^{-}\{- \}) & (2) \\
 \Rightarrow & *C(.CH_{-}\{3\})_{-}\{2\}(.NN^{+}\{+ \}N^{-}\{- \}) & (3) \\
 \Rightarrow & *C(.CHHH)(.CHHH)(.NN^{+}\{+ \}N^{-}\{- \}) & (4) \\
 \Rightarrow & *?C(-. ?C?H?H?H)-(-. ?C?H?H?H) & \\
 & -(-. ?N?N^{+}\{+ \}N^{-}\{- \}) & (5)
 \end{aligned}$$

式 (1) が入力となる示性式である。“*”は化学構造型の主構造への接続先を表す記号である。“{ }”は下付き数字を表す記述方法である。示性式中ではメチル基 “Me” やアジ基 “N₃” などの基ないし原子団の略称が頻繁に用いられる。それらが示す分子構造を得るために、事前に略称を登録したテーブルを使用して式 (2) のような文字列への変換を行う。ここで使用するテーブルとは、有限個の基ないし原子団の略称を扱うものであり、従来の化学構造型 OCR アプリケーションが使用する辞書とは意味が異なる。“{+}”は正電荷を表す記述方法、“{-}”は負電荷を表す記述方法である。分岐または繰返しを意味する括弧が存在する場合は、式 (3) のように括弧内の先頭に括弧外への接続があることを意味する記号 “.” を挿入する。括弧内での原子の手の数の計算をする際、括弧外との結合で使用する手の数を先に除いて計算を行うためにこのような記号を用いる。原子記号や括弧の後にその個数を表す数字がある場合は、式 (4) のように数字に合わせた複製を行う。原子記号や括弧の間に、“-”, “=”, “≡” などの結合の種類を指定する記号がない箇所には、式 (5) のように結合の種類が未知であることを意味する結合記号 “?” を挿入する。また、括弧で括られた原子団は、通常は他の原子と単結合するため、開き括弧の前後に単結合を表す記号 “-” を挿入する。2.2 節で述べたように、化学構造型の主構造の右側に描かれた示性式の場合は、表記の方向の異なる3パターンに合わせた3つの展開文字列が生成される。

4.2 展開文字列を生成する形式文法

展開文字列の構文を記述するために適した形式文法を考える。高速な処理を行うためには正規文法や LALR (1) 文法、LR (1) 文法のような決定性文法を用いたい。しかしながら、示性式には括弧の入れ子構造が存在するため正規文法では記述不能である。また、示性式には複数の手の数をとる原子によって生じる曖昧性が存在するため (2.2 節図 4 参照)、曖昧性を許さない LALR (1) 文法および LR (1) 文法では記述することができない。したがって、提案手法では文脈自由文法を使用する。文脈自由文法の構文解析も多項式時間 (文字列の長さを n とすると、 $O(n^3)$ 時間) でできるため、十分高速である。

展開文字列の構文を定義する文法を図 9 に示す。図 9 (1) の “SS” は開始記号である。“A₀”, “A₁”, “A₂”, “A₃”, “A₄”, “A₅”, “A₆” は非終端記号である。ダブルクォーテーションで括られている “H”, “Tos”, “?”, “=”, “(”, “)” などの文字列は終端記号である。

非終端記号の下付き数字は、他の原子と結合するために使用可能な手の数を表している。たとえば、水素原子の手の数は1であるから、手の数が1であることを表す非終端記号 “A₁” から水素原子を表す終端記号 “H” を生成する規則 $A_1 \rightarrow \text{“H”}$ が対応する。図 9 (16)~(23) は、原子ないし原子団を表す非終端記号からそれぞれの手の数に合った非終端記号に変換する生成規則である。

一方、図 9 (2)~(10) は、非終端記号で表されている原子どうしの結合を表す生成規則である。たとえば、手の数が1の原子 “A₁” と2の原子 “A₂” が結合する場合は、両者が手の数を1本ずつ出し合う。すると手の数は前者が0、後者が1となるため、手の数が1の原子を表す “A₁” から生成する規則 $A_1 \rightarrow A_1 \text{“?”} A_2$ が対応する。もし結合の種類が単結合を表す結合記号 “-” で指定されている場合は $A_1 \rightarrow A_1 \text{“-”} A_2$ が対応する。

また、図 9 (11)~(16) は分岐または繰返しを表す括弧で括られた原子団を変換するための生成規則である。括弧内の手の数の計算では、括弧外との結合で使用する手の数を先に除いて計算を行っている。そのため、括弧内の手の数の計算が正常に行われた後に、先に除いた手の数を戻す必要がある。たとえば、括弧内の先頭の原子が括弧外の原子と単結合している場合は、括弧内の先頭の原子の前に手の数が1本の原子を表す記号を加える。括弧内の手の数の計算が正常に行われ、括弧内が手の数が余っていないことを表す非終端記号 “A₀” になったとすると、先に除いた手の数1本が戻されるため、“A₁” から生成する規則 $A_1 \rightarrow \text{“(”} A_0 \text{“)”}$ が対応する。

4.3 構文解析

展開文字列はこの文法によって構文解析され、構文木が生成される。例として、図 9 の文法からカルボキシ基の展

- (1) $SS \rightarrow A_0$
- (2) $A_0 \rightarrow A_1 \text{ “?” } A_1 \mid A_2 \text{ “?” } A_2 \mid A_3 \text{ “?” } A_3 \mid A_1 \text{ “-” } A_1 \mid A_2 \text{ “=” } A_2 \mid A_3 \text{ “#” } A_3$
- (3) $A_1 \rightarrow A_1 \text{ “?” } A_2 \mid A_2 \text{ “?” } A_1 \mid A_2 \text{ “?” } A_3 \mid A_3 \text{ “?” } A_2 \mid A_3 \text{ “?” } A_4 \mid A_4 \text{ “?” } A_3$
- (4) $\mid A_1 \text{ “-” } A_2 \mid A_2 \text{ “-” } A_1 \mid A_2 \text{ “=” } A_3 \mid A_3 \text{ “=” } A_2 \mid A_3 \text{ “#” } A_4 \mid A_4 \text{ “#” } A_3$
- (5) $A_2 \rightarrow A_1 \text{ “?” } A_3 \mid A_3 \text{ “?” } A_1 \mid A_2 \text{ “?” } A_4 \mid A_4 \text{ “?” } A_2 \mid A_3 \text{ “?” } A_5 \mid A_5 \text{ “?” } A_3$
- (6) $\mid A_1 \text{ “-” } A_3 \mid A_3 \text{ “-” } A_1 \mid A_2 \text{ “=” } A_4 \mid A_4 \text{ “=” } A_2 \mid A_3 \text{ “#” } A_5 \mid A_5 \text{ “#” } A_3$
- (7) $A_3 \rightarrow A_1 \text{ “?” } A_4 \mid A_4 \text{ “?” } A_1 \mid A_2 \text{ “?” } A_5 \mid A_5 \text{ “?” } A_2 \mid A_3 \text{ “?” } A_6 \mid A_6 \text{ “?” } A_3$
- (8) $\mid A_1 \text{ “-” } A_4 \mid A_4 \text{ “-” } A_1 \mid A_2 \text{ “=” } A_5 \mid A_5 \text{ “=” } A_2 \mid A_3 \text{ “#” } A_6 \mid A_6 \text{ “#” } A_3$
- (9) $A_4 \rightarrow A_1 \text{ “?” } A_5 \mid A_5 \text{ “?” } A_1 \mid A_2 \text{ “?” } A_6 \mid A_6 \text{ “?” } A_2 \mid A_1 \text{ “-” } A_5 \mid A_5 \text{ “-” } A_1 \mid A_2 \text{ “=” } A_6 \mid A_6 \text{ “=” } A_2$
- (10) $A_5 \rightarrow A_1 \text{ “?” } A_6 \mid A_6 \text{ “?” } A_1 \mid A_1 \text{ “-” } A_6 \mid A_6 \text{ “-” } A_1$
- (11) $A_1 \rightarrow \text{ “(-” } A_0 \text{ “)” } ”$
- (12) $A_2 \rightarrow \text{ “(-” } A_1 \text{ “)” } \mid \text{ “(=” } A_0 \text{ “)” } ”$
- (13) $A_3 \rightarrow \text{ “(#” } A_0 \text{ “)” } ”$
- (14) $A_4 \rightarrow \text{ “(=” } A_2 \text{ “)” } ”$
- (15) $A_6 \rightarrow \text{ “(#” } A_3 \text{ “)” } ”$
- (16) $A_0 \rightarrow \text{ “H” } \{ + \} \mid \text{ “Na” } \{ + \} \mid \text{ “K” } \{ + \} \mid \text{ “Al” } \{ 3+ \} \mid \text{ “Ca” } \{ 2+ \} \mid \text{ “Zn” } \{ 2+ \} \mid \text{ “Gd” } \{ 3+ \}$
- (17) $\mid \text{ “Cl” } \{ - \} \mid \text{ “Br” } \{ - \} \mid \text{ “I” } \{ - \}$
- (18) $A_1 \rightarrow \text{ “H” } \mid \text{ “F” } \mid \text{ “Na” } \mid \text{ “Cl” } \mid \text{ “K” } \mid \text{ “Br” } \mid \text{ “Ag” } \mid \text{ “I” } \mid \text{ “O” } \{ - \} \mid \text{ “R” } \mid \text{ “Ph” } \mid \text{ “*” } \mid \text{ “.” } ”$
- (19) $A_2 \rightarrow \text{ “O” } \mid \text{ “Mg” } \mid \text{ “S” } \mid \text{ “Ca” } \mid \text{ “Zn” } \mid \text{ “Hg” } \mid \text{ “Co” } \{ + \} \mid \text{ “Pt” } \{ 2+ \} \mid \text{ “N” } \{ - \} \mid \text{ “Pph” } \mid \text{ “$” } \mid \text{ “:” } ”$
- (20) $A_3 \rightarrow \text{ “N” } \mid \text{ “Al” } \mid \text{ “P” } \mid \text{ “Co” } \mid \text{ “As” } \mid \text{ “Gd” } \mid \text{ “Bi” } \mid \text{ “S” } \{ + \} \mid \text{ “%” } \mid \text{ “;” } ”$
- (21) $A_4 \rightarrow \text{ “C” } \mid \text{ “Si” } \mid \text{ “S” } \mid \text{ “Ge” } \mid \text{ “Pt” } \mid \text{ “N” } \{ + \} ”$
- (22) $A_5 \rightarrow \text{ “P” } \mid \text{ “Sb” } ”$
- (23) $A_6 \rightarrow \text{ “S” } \mid \text{ “Tc” } ”$

図 9 展開文字列の構文を定義する文法

Fig. 9 The grammar that defines syntax of unfolded strings.

開文字列 “*?C?O?O?H” を生成する様子を以下に示す。

$SS \Rightarrow A_0$
 $\Rightarrow A_1 ? A_1$
 $\Rightarrow A_1 ? H$
 $\Rightarrow A_1 ? A_2 ? H$
 $\Rightarrow A_1 ? O ? H$
 $\Rightarrow A_3 ? A_2 ? O ? H$
 $\Rightarrow A_3 ? O ? O ? H$
 $\Rightarrow A_1 ? A_4 ? O ? O ? H$
 $\Rightarrow A_1 ? C ? O ? O ? H$
 $\Rightarrow * ? C ? O ? O ? H$

結果として得られた構文木を図 10 に示す。

4.4 完全構造式の生成

構文解析によって得られた構文木から、完全構造式を生成するために必要な原子間の結合情報を得ることができる。図 10 を例として説明する。

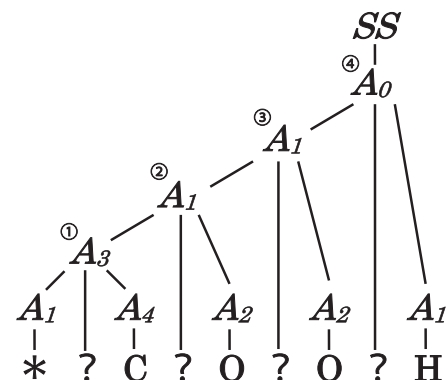


図 10 “*?C?O?O?H” の構文木

Fig. 10 Parse tree of “*?C?O?O?H”.

まず初めに、図 10 の ① を根とする部分木に注目すると、生成規則 $A_3 \rightarrow A_1 \text{ “?” } A_4$ が適用されていることが分かる。これは、手の数が 4 本の炭素原子 “C” と、化学構造式の主構造への接続部を表す記号で手の数が 1 本の “*” が単結合し、炭素原子 “C” の手の数が 3 本余っていることを意味する。次に、図 10 の ② を根とする部分木に注目すると、生成規則 $A_1 \rightarrow A_3 \text{ “?” } A_2$ が適用されていることが

分かる。これは、余っている手の数が1本の炭素原子“C”と、2本の手の数を持つ酸素原子“O”が二重結合し、炭素原子“C”の手の数が1本余っていることを意味する。次に、図10の③を根とする部分木に注目すると、生成規則 $A_1 \rightarrow A_1 \text{ “?” } A_2$ が適用されていることが分かる。これは、余っている手の数が1本の炭素原子“C”と、2本の手の数を持つ酸素原子“O”が単結合し、酸素原子“O”の手の数が1本余っていることを意味する。次に、図10の④を根とする部分木に注目すると、生成規則 $A_0 \rightarrow A_1 \text{ “?” } A_1$ が適用されていることが分かる。これは、余っている手の数が1本の酸素原子“O”と、1本の手の数を持つ水素原子“H”が単結合し、どの原子の手の数も余っていないことを意味する。

こうして、構文木から得られた原子間の結合情報を記録していくことで、最終的に完全構造式を生成することができる。

5. 評価実験

提案手法を評価するために、C言語を用いて、与えられた示性式から展開文字列を生成するプログラムと図9の文脈自由文法によって展開文字列を構文解析するプログラムを作成した。自由文脈文法の構文解析アルゴリズムとして、CYK法[10]を使用した。化学構造式画像から示性式を抽出する作業には、著者らが新しく開発中である化学構造式OCRアプリケーションを使用する（文字と線が重なっている画像、低解像度の画像など、示性式の抽出がうまくいかない画像には、人手で修正することを許す）。

作成したプログラムを用いて、JAPICの日本の医薬品構造式集[9]の中の高分子を取り除いた1,427個の化学構造式にテストを行った（高分子はアミノ酸のペプチド配列であり、示性式が使われていないため除外した）。その結果、94%（1,346/1,427）の化学構造式に登場する示性式を正確に認識できたことを確認した。図11(a)は正確に認識された示性式の例である。主構造への接続部分に*を加え、展開文字列に変換すると“*?C?H?H?O?S?O?O?Na”になる。この展開文字列を構文解析し、得られた完全構造式を図11(b)に示す。

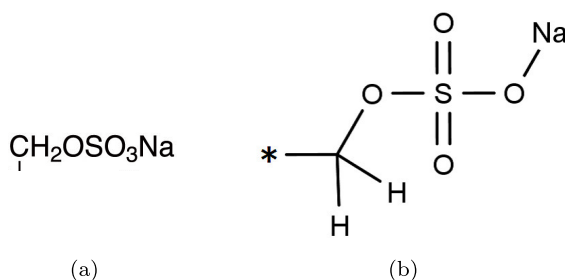


図11 (a) 正しく認識された示性式と (b) 得られた完全構造式
Fig. 11 (a) Condensed structural formula recognised correctly and (b) complete structural formula obtained.

正しく認識することができなかった81個の示性式は、以下の3つに分類される。

- (1) 示性式とは異なる表記方法の文字列が存在する場合。(71個)
- (2) 示性式が示す部分構造中に共有結合とは異なるイオン結合や配位結合が含まれている場合。(1個)
- (3) 原子を表す記号に同位体を区別するための質量数を表す数字が添えられている場合。(9個)

(1)の例を図12に示す。図12の矢印が指す文字列は、分子式となっている。分子式は示性式と同様にその化合物を構成する元素を表す記号を用いた文字列であるが、原子の種類とそれぞれの原子の量を簡潔に表現した式であり、示性式のように部分構造が明確ではない。分子式を完全構造式に変換するには、より多くの化学の知識を必要とする。我々の提案手法は入力として示性式の表記方法だけを想定しているため、示性式以外の文字列は処理できなかった。

(2)の例を図13に示す。図13の矢印が指す示性式には、イオン結合が含まれている。我々の提案手法で使用する文法は共有結合を想定した原子価の計算を基に定義しているため、共有結合とは異なる結合の種類を含む一部の示性式は処理できなかった。

(3)の例を図14に示す。図14の矢印が指す原子記号には、その原子の質量数を表す数字が添えられている。こ

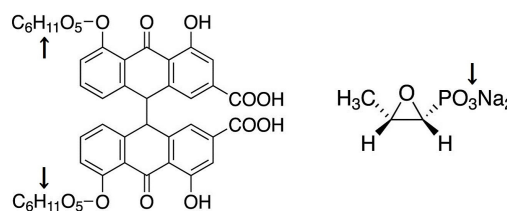


図12 分子式（非示性式）

Fig. 12 Molecular formulas (they are not a condensed structural formula).

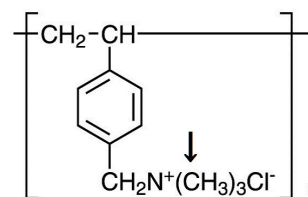


図13 イオン結合を含む示性式

Fig. 13 Condensed structural formulas with ionic bonds.

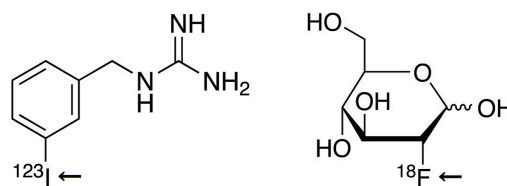


図14 質量数をともなう原子

Fig. 14 Atoms with atomic number.

れは同じ原子番号を持つが中性子の数が異なる同位体を区別するための表記方法である。我々の提案手法では同位体の表記を扱っていないため、同位体を含む文字列は処理することができなかった。

6. まとめ

本稿では、形式文法を用いて化学構造式中の示性式を自動認識する方法を提案した。JAPICの日本の医薬品構造式集 [9] の中から高分子を取り除いた 1,427 個の化学構造式を用いてテストを行った。その結果、94% (1,346/1,427) の化学構造式に登場する示性式を正確に認識できた。これより、本提案手法はかなり実用性が高いと考える。

今後の課題として、まだ正確に認識することができていない 3 つのケースに対応していくことがあげられる。イオン結合に関しては、本研究で定義した文法を拡張して対応することが期待できる。分子式から完全構造式への変換に関しては、先行研究などの調査も含めて始めていきたいと考えている。また、同素体などの質量数を表す数字に関しては、アドホックに対応していけばよいと考えている。

また、化学構造式全体の認識を文法処理として行うことにも取り組みたいと考えている。そのためには、依存構造解析によって原子間の結合を依存関係として取得する手法や、既存の化学構造を生成する文法など [11], [12] を参考に、化学構造式全体を文法によって再定義する必要がある。

謝辞 本研究は JSPS 科研費 26282044 の助成を受けたものである。

参考文献

- [1] Sadawi, N.: A rule-based approach for recognition of chemical structure diagrams, Ph.D. thesis, University of Birmingham (2013), available from (<http://etheses.bham.ac.uk/4325/>).
- [2] Sadawi, N., Sexton, A. and Sorge, V.: MolRec at CLEF 2012-Overview and Analysis of Results, *CLEF 2012 Chemical Structure Recognition Task* (2012).
- [3] Sadawi, N., Sexton, A. and Sorge, V.: Performance of MolRec at TREC 2011 Overview and Analysis of Results, *Text REtrieval Conference (TREC 2011)* (2011).
- [4] Sadawi, N.: Recognising chemical formulas from molecule depictions, *Pre-Proc. 8th IAPR International Workshop on Graphics Recognition (GREC 2009)*, pp.167–175 (2009).
- [5] Filippov, I.V. and Nicklaus, M.C.: Extracting chemical structure information: Optical structure recognition application, *Pre-Proc. 8th IAPR International Workshop on Graphics Recognition (GREC 2009)*, pp.133–142 (2009).
- [6] Fujiyoshi, A., Nakagawa, K. and Suzuki, M.: Robust Method of Segmentation and Recognition of Chemical Structure Images in ChemInfty, *Pre-Proc. 9th IAPR International Workshop on Graphics Recognition (GREC 2011)*, pp.121–125 (2011).
- [7] Karzel, D., Nakagawa, K., Fujiyoshi, A. and Suzuki, M.: Inconsistency-Driven Chemical Graph Construction in ChemInfty, *Proc. 9th IAPR International Workshop on Graphics Recognition (GREC 2011)*, pp.119–128 (2010).
- [8] Nakagawa, K., Fujiyoshi, A. and Suzuki, M.: Ground-truthed dataset of chemical structure images in Japanese published patent applications, *Proc. 9th IAPR International Workshop on Document Analysis Systems (DAS2010)*, pp.455–462 (2010).
- [9] 一般財団法人日本医薬情報センター (JAPIC): 日本の医薬品構造式集, 丸善出版株式会社 (2010).
- [10] Hopcroft, J.E. and Ullman, J.D. (著), 野崎昭弘, 木村泉 (訳): 言語理論とオートマトン, サイエンス社 (1971).
- [11] Ehlers, E.M. and von Solms, S.H.: Using random context structure grammars to represent chemical structures, *Information Processing Letters*, Vol.30, No.3, pp.159–166 (1989).
- [12] Proschak, E., Wegner, J.K., Schüller, A., Schneider, G. and Fechner, U.: Molecular query language (MQL)—A context-free grammar for substructure matching, *J. Chem. Inf. Model.*, Vol.47, No.2, pp.295–301 (2007).



佐藤 文恭

1991 年生。2014 年茨城大学工学部情報工学科卒業。2016 年同大学大学院理工学研究科情報工学専攻博士前期課程修了。現在、株式会社日立ハイテクノロジーズ勤務。



藤芳 明生 (正会員)

1971 年生。1997 年電気通信大学大学院博士前期課程修了。2000 年同博士後期課程修了。博士(理学)。2000 年茨城大学助手。2005 年同大学講師。2011 年同大学准教授。形式言語理論の研究に従事。電子情報通信学会会員。