Окрім того, MATLAB містить пакет **Simulink**, що дозволяє проводити моделювання об'єктів, процесів і систем керування шляхом опису логіки їх функціонування у вигляді структурних схем, а не аналітичних залежностей.

Такий широкий набір можливостей разом з відносною простотою виконання основних процедур програмування, моделювання та візуального оформлення результатів дозволяє вважати програмний пакет MATLAB найбільш ефективним засобом для дослідження і проектування різноманітних технологічних об'єктів, процесів і систем автоматичного керування в поліграфічному машинобудуванні.

1. Гультяев А.К. MatLAB 5.2. Имитационное моделирование в среде Windows: Практическое пособие. СПб., 1999. 2. Заднішевський О.Ю., Петрів Р.І. Алгоритм комп'ютерного моделювання технологічних об'єктів у середовищі MATLAB/SIMULINK // Комп'ютерні технології друкарства: Зб. наук. пр. Львів, 2000. № 5. 3. Лазарев Ю.Ф. MatLAB 5.x. К., 2000.

УДК 004.22+544.1

АНАЛІЗ ТА КЛАСИФІКАЦІЯ СИСТЕМ КОДУВАННЯ ХІМІЧНИХ ФОРМУЛ

Т.В. Нерода

Формулюються основні вимоги до систем кодування зображень хімічних структур для використання в комп'ютерно-видавничих системах. Пропонується оригінальна класифікація систем кодування хімічних формул.

Формулируются основные требования к системам кодирования изображений химических структур для использования в компьютерно-издательских системах. Предлагается оригинальная классификация систем кодирования химических формул.

Для набору у комп'ютерно-видавничих системах (КВС) зображень хімічних структур з метою подальшого збереження та редагування необхідно передусім подати формулу у вигляді лінійної множини символів (лінійного запису, коду), прийнятного для машинної інтерпретації, тобто такою, що припускає формалізований (алгоритмічний) аналіз. Сукупність граматичних правил і відповідної лексики, яка дозволяє перетворити побудоване графічне зображення формули, зокрема структурної, та особливості розташування її фрагментів у лінійний запис, ставить у відповідність кожній формулі слово в заданому алфавіті, називається вхідною мовою або системою кодування (СК) хімічних сполук [6].

Конкретний клас задач, вирішуваних відповідною системою кодування, обумовлює свої вимоги до неї. Таким чином, системи кодування хімічної інформації можна оцінювати за такими критеріями:

відсутність втрат і надлишку інформації при кодуванні;

відсутність надлишкової інформації у вигляді окремих доповняльних процедур, таблиць відповідності та номенклатурних баз даних;

унікальність лінійного запису для вихідного зображення;

охоплення значної кількості сполук, фрагментів сполук або пристосовуваність до окремих класів сполук;

збережуваність графічної просторової організації вихідного зображення.

Відповідно до застосовуваних методів кодування формул системи кодування хімічної інформації можна класифікувати як системи з повним і неповним збереженням структурних фрагментів (див. таблицю). Системи з повним збереженням і наступним відображенням структурно-хімічної інформації, у свою чергу, діляться за ступенем деталізації фрагментів формули, які позначаються окремим символом лінійного запису, на поатомні, дрібноблокові, крупноблокові.

Системи кодування хімічних формул

Системи кодування хімічних формул				Критерії										
				Відсутність втрат інформації	Унікальність коду	Універсальність			Відсутність надлишкової інформації			Використання кодів ANSI	Збережуваність геометрично-позиційних х-к	Врахування технологічних вимог набору
						лінійні	цикли	зв'язки	запис	процед.	БД			
Системи кодування з повним відображенням структурних фрагментів	Поатомні (топологічні) системи	матриці зв'язків	Спіалтера	+	+	+	+	+	+	–	+	+	–	–
			Глюка	+	+	+	+	+	+	–	+	+	–	–
		лінійні	Балларда	+	+	+	+	+	–	–	+	+	–	–
			Хва-Айсмана	+	+	+	–	+	+	–	+	+	–	–
		геометричні	Вальдо	+	+	+	+	+	+	–	–	+	+	–
			HEXAGON	+	+	+	+	+	+	–	+	+	+	–
			Майєра	+	+	+	+	+	–	–	+	+	+	–
			Пантюхіної	+	+	+	+	+	+	–	–	–	+	–
	Дрібноблокові		Locus	+	–	+	+	+	+	–	–	+	–	–
			Лефковіча	+	–	+	–	+	–	–	–	–	–	–
			Хейварда	+	–	+	+	–	–	–	–	+	–	–
			'вузли-шляхи'	+	–	+	–	+	+	–	–	+	–	–
	Крупноблокові	контурно-вузлові	Вісвессера	+	+	+	+	+	+	–	–	+	–	–
			Дайсона	+	+	+	+	+	–	–	–	–	–	–
		фрагментарні	Д Б	+	+	+	+	+	–	–	–	–	–	–
			П Н К	+	+	+	+	+	+	+	–	–	+	–
Неповні	Дескриптори		GREMAS	–	–	–	+	–	+	–	–	–	–	–
			Шевякової	–	–	+	+	–	+	–	–	–	–	–
			NRC	–	–	–	+	–	+	–	–	–	–	–
	Зв'язні		Пенні	–	–	+	+	–	+	–	–	+	–	–

Поатомні (топологічні) системи кодування передбачають відповідність кожному атому структурної формули окремого символу коду. Зважаючи на явну недоцільність канонічного поатомного запису складних органічних молекул, усі відомі поатомні системи кодування є неканонічними. Поатомні коди структурної формули, як правило, являють собою ту чи іншу модифікацію матриці інциденцій молекулярного графу, відповідного структурній формулі з довільно пронумерованими атомами: у матриці і-й вершині графа ставиться у відповідність і-й рядок та і-й стовпець й одиниці проставляються на перетині рядка і стовпця, відповідно до з'єднаних вершин [3,10].

Така довільна нумерація не сприяє збереженню позиційних характеристик елементів зображення хімічної сполуки. У разі подання елементів з постійною валентністю матриця підлягає скороченню шляхом видалення водневих атомів, відновлення яких можливе з використ-

танням окремого процедурного алгоритму, що значно ускладнює програму кодування. Деякі поатомні СК передбачають лінійний запис вхідного коду з поданням додаткових відомостей про загальну кількість циклів й ациклічних ланцюгів, в які входить даний елемент, і водневих атомів, інцидентних з ним [8]; ця інформація суттєво збільшує обсяг запису, що також ускладнює алгоритм кодування.

При кодуванні за поатомними системами, які в певній мірі зберігають геометричні характеристики вхідного зображення, символи зв'язків і вершин записуються в послідовній відповідності до їх розташування у вихідній структурній формулі [1, 3, 5]. Проте подальша трансформація запису в табличне подання також веде до надлишковості інформації та громіздкості коду.

У дрібноблокових системах кодування невелика група атомів позначається окремим символом конкретно визначеного алфавіту [2, 3, 7]. Шифрування сполуки дрібноблоковими системами кодування зводиться до перерахунку довільно пронумерованих вузлових атомів, що формують вуглецевий скелет, блоків і зв'язків між ними. При такому записі виникає ситуація, коли одному коду можуть відповідати декілька зображень. До недоліків систем можна також віднести наявність таблиці відповідності у вигляді бази даних, складних правил кодування та принципове долучення до запису надлишкової інформації про спеціальні атоми (масове число, заряд, валентність стереоізомери тощо).

Крупноблокові системи кодування застосовуються для відображення великих фрагментів (блоків) найпоширеніших структур у вигляді одного символу: зокрема, передбачено особливе односимвольне подання атомних груп, кілець, галогенів, розгалужених атомів азоту та вуглецю. Для їх наступного розшифрування застосовується спеціальна процедура із звертанням до номенклатурної бази даних. СК залежно від логічного принципу кодування циклічних систем органічних сполук, зокрема, поділяються на фрагментарні і контурно-вузлові. Контурно-вузлові системи використовують поняття замкненого контуру з описом зв'язків між вузлами контуру або кількості атомів у кільцях, в які входить даний вузол. Передбачено ряд правил старшинства з врахуванням максимальної кількості розгалужених атомів, наявності та типу гетероатомів, локантів, насиченості та деяких інших чинників [1, 3]. Загалом контурно-вузловим системам властивий суттєвий недолік усіх канонічних СК. Поряд із значною кількістю складних правил старшинства та використанням символіки, непридатної для застосування в машинних системах кодування, вони не забезпечують однозначного відтворення хімічної структури, втрачаючи певні інформаційні характеристики зображення формули.

Фрагментарна система кодування передбачає декомпозицію циклічного графа на складові блоки відповідно до заданого списку блоків [3, 5]. Для однієї і тієї ж структурної формули можливі різні варіанти розбиття на блоки з наступною довільною нумерацією їх та їхніх вершин. Код циклічної сполуки містить у загальному випадку інформацію про усі виділені циклічні блоки і вільні замісники (ациклічні фрагменти), організацію з'єднання циклічних блоків, спосіб зв'язку між атомами сполуки. Ациклічний фрагмент записується як перелік груп атомів головного ланцюга; відгалужена ланка зазначається в дужках після вузлового атома. При зміні вибору головного ланцюга або при іншому варіанті розбиття на блоки мають місце втрати інформації про розташування атомів у молекулі. Крім того, за наявності відхилення від стандартного виду блоків у декількох фрагментах факт відхилення зазначається в коді лише для одного фрагменту, що може спричинити виникнення помилок при декодуванні запису.

Використання заданого списку блоків, котрий значно зменшує гнучкість системи, збільшує обсяг і складність програми кодування, можна уникнути шляхом введення алгоритму, що дозволяє елементарно кодувати декілька тисяч поліциклічних структур, до яких просто не зводяться підструктури описаного списку [4]. Базовими структурами в таких системах з поблочною нумерацією кілець вважаються регулярні циклічні структури з шестичленними кільцями; вершини кільця позначаються певними літерами за годинниковою стрілкою, починаючи з верхньої вершини. Горизонтальний ряд кілець регулярної структури, що містить кільце, розташоване лівіше за решту, нумерується цифрами непарного ряду; кільцям суміжних горизонтальних рядів присвоюються парні номери. Таким чином, реалізується позиційний принцип кодування скелета регулярних структур, за яким досить легко ідентифікувати задану вершину певного кі-

льця. При кодуванні ациклічних ланцюгів лінійний запис формується послідовним перерахуванням скорочень атомних груп головного ланцюга з кодуванням кратних зв'язків; відгалуження ієрархічно фіксується відповідно до вузла. Таке збереження в коді вихідної організації зображення в досить явному вигляді дозволяє застосовувати нескладні алгоритми для машинного опрацювання лінійного запису з наступним коректним відтворенням зображення. Проте очевидним недоліком таких систем є відсутність можливості подання структури кільцями, зображеними правильними n-кутниками (n<>6), та ускладнене кодування квазірегулярних структур, що потребує наявності людського фактора.

Відповідно до прийнятої термінології усі розглянуті системи кодування названі системами з повним відображенням структурної хімічної інформації. Проте для розв'язання ряду часткових інформаційних задач використовуються системи кодування із запрограмованою втратою інформації, так звані неповні системи кодування. Найпоширенішим класом неповних систем кодування є клас фрагментарних мов дескрипторного типу. При цьому задається деякий список дескрипторів, що відображають загальні відомості про сполуку (тип циклічної системи, тип гетероатомів, фрагментів замісників, зв'язок тощо), і код сполук являє собою список номерів дескрипторів [3, 9].

Інший метод формування запису застосовується у зв'язних неповних системах кодування [11]. Зв'язний код переводить у лінійний запис графічну структуру навколо індивідуального атома за допомогою трьох рівнів зв'язності. Цей код не відображає єдиного унікального просторового розташування зв'язків, але має загальний характер, незалежно від того, як обертається чи інвертується структура.

У цілому неповні системи кодування з обов'язковим членуванням на фрагменти та реєстрацією їх входження не дозволяють у повній мірі реалізувати лінійний запис із наступним збереженням структурно-позиційної інформації про сполуку, оскільки одна з основних умов формування зображення – врахування взаєморозташування та взаємозв'язку окремих фрагментів – недостатньо відображена в коді.

Описаний аналіз систем кодування свідчить, що відомі методи кодування не в повній мірі підходять для застосування їх у комп'ютерних видавничих системах, тому виникає завдання створення нових або суттєва модифікація наявних методів кодування сполук. Сукупність лексики та граматичних правил такої системи кодування як інструменту КВС не повинна передбачати можливості виявлення помилок, викликаних порушенням валентності і хибним поданням символів елементів, містити надлишкову інформацію про склад сполуки та скорочене зображення атомних груп символозамінниками тощо, що пов'язано з наявністю деталізованих таблиць відповідності та громіздкої номенклатурної бази даних. Відтак запис хімічної структури грунтуватиметься на простих правилах ідентифікації ознак кожного елемента зображення відповідно до окремого символу коду із збереженням геометрично-позиційних характеристик і виконанням технологічних вимог правильного поліграфічного відтворення й буде придатний для реалізації підпрограми формування зображень хімічних структур у комп'ютерних видавничих системах.

1. Влэдуц Г.Э., Гейвандов Э.А. Автоматизированные информационные системы для химии. М., 1973. 2. Влэдуц Г.Э., Стоцкий Э.Д., Финн В.К. О некоторых системах записи структурных формул органической химии // Доклады конференции по обработке информации, машинному переводу и автоматизированному чтению текста. Вып. 4. М., 1961. 3. Вопросы разработки механизированной информационно-поисковой системы для центрального справочного информационного фонда по химии и химической промышленности. Разработка методики прямого кодирования химических соединений. Вып. 10. М., 1968. 4. Гейвандов Э.А. Разработка информационно-поисковой системы для соединений с сопряженными связями: Автореф. дис. к.х.н. М., 1970. 5. Пантюхина М.Е. Ввод информации о структуре химического соединения в машину и вывод из машины // Доклады конференции по обработке информации, машинному переводу и автоматизированному чтению текста. Вып. 1. М.,1961. 6. Першиков В.И., Савинков В.М. Толковый словарь по информатике. М., 1991. 7. Bouman H. Linearly organized chemical code for use in computer systems (Locus). J. Chem. Doc., 1963, 3, № 2. С. 92. 8. Eismann S.H. A Polish-type notation of chemical structures. J. Chem. Doc., 1964, 4, № 3. С. 186–190. 9. Fugman R., Braun W., Vaupel W. GREMAS – Ein Weg zur Klassificatin und und Documentation in der organishen Chemie. Nachr. Doc., 1963, 14, № 4. С. 179. 10. Gluck D.J. A chemical structures storage and search system developed at Du Pont. J. Chem. Doc., 1965, 5, № 1. С. 43. 11. Penny R.H. A connectivity code for use in describing chemical structures. J. Chem. Doc., 1965, 5, № 2. С. 113.