Representação gráfica de moléculas na reconstrução molecular de misturas de hidrocarbonetos

Aluno: Lucas Feliciano da Silva
Orientador: Prof. Dr. Dirceu Noriler
Laboratório de Pesquisa em Processos Químicos e Gestão Empresarial (PQGe)
Faculdade de Engenharia Química (FEQ), UNICAMP

Palavras-chave

Structure Oriented Lumping
SMILES
Representação gráfica de moléculas

RESUMO

Um algoritmo para representação gráfica de moléculas encontra-se em desenvolvimento para auxiliar a visualização dos resultados obtidos a partir de reconstrução molecular de misturas de petróleo no método SOL.

Introdução

Com a disponibilidade de misturas de petróleo pesadas e a demanda por combustíveis mais leves, a análise de cargas por métodos de reconstrução molecular tornou-se importante nos estudos dos processos de refino do petróleo. Apesar da grande complexidade e alto número de diferentes componentes presentes nas misturas de hidrocarbonetos pesados, os dados que atualmente podem ser obtidos não são capazes de representar analiticamente todos esses componentes [1]. Faz-se, então, necessário o uso de técnicas que representem de modo sucinto múltiplas possíveis moléculas para avaliação de rotas cinéticas e produtos possíveis em condições definidas.

Nesse contexto, o método *SOL* (*Structure Oriented Lumping* ou Agrupamento Orientado à Estrutura) tem sido amplamente utilizado na indústria de petróleo[1] e tem por base a representação de moléculas na forma vetores, onde cada posição corresponde a uma estrutura básica e o número presente nessas posições representa a quantidade da estrutura correspondente presente na molécula. Apesar de ser eficiente na reconstrução molecular de cargas e produtos do processo, a representação em forma de vetores pode não ser muito clara, pois requer o conhecimento exato de qual estrutura corresponde a qual posição do vetor. Ademais, um vetor pode representar várias moléculas e essa ambiguidade dificulta a visualização das moléculas, principalmente em situações como a de representação de uma reação, onde seus reagentes e produtos necessitam ser bem definidos.

Portanto, é conveniente utilizar programas que permitam representar de forma gráfica as informações obtidas pelo método SOL. Atualmente existem vários formatos de arquivo e notações que podem ser utilizados para denotar moléculas e a partir deles gerar as ilustrações gráficas. Dentre esses, destaca-se a notação SMILES, devido a sua fácil compreensibilidade e devido ao seu amplo suporte por programas e bibliotecas na área da quimio-informática. SMILES é um sistema que pode representar moléculas através de uma notação linear de caracteres que represente os grafos correspondente às fórmulas esqueléticas das moléculas [2].

O objetivo desse projeto é auxiliar a visualização das moléculas reconstruídas pelo método SOL através de representações gráficas. Com essa finalidade, está sendo desenvolvido um programa que aceite por entrada os dados adquiridos pela reconstrução molecular e que apresente de forma clara e rápida as fórmulas esqueléticas das moléculas reconstruídas.

Método SOL

No método de reconstrução SOL, cada molécula pode ser representada por um vetor onde as posições correspondem aos principais grupos encontrados na maioria das moléculas de hidrocarbonetos pesados. O número de estruturas (posições) presentes no vetor pode variar dependendo dos critérios e objetivos do uso

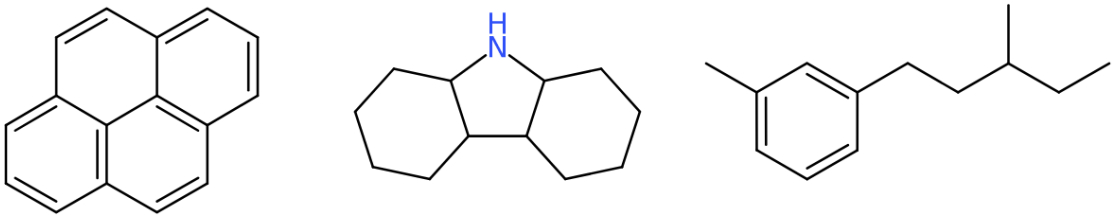
do método e da sua aplicação, entretanto essas estruturas não se diferenciam muito do método SOL originalmente convencionado, onde o vetor SOL possui 22 posições e a ordem de reconstrução da molécula é feita de forma incremental, salvas exceções de posições que não representam estruturas como ligações ou insaturações das estruturas[3]. A tabela 1 apresenta os atributos usualmente presentes em um vetor SOL.

Tabela 1: Vetor SOL

A6	A4	A2	N6	N5	N4	N3	N2	N1	R	br	me	IH	AA	NS	RS	AN	NN	RN	NO	RO	KO

As 3 primeiras posições, A6, A4 e A2, correspondem às estruturas de anéis aromáticos. Enquanto a primeira posição, A6, denota um anel aromático principal, as posições A4 e A2 denotam anéis aromáticos que se ligam a outros anéis, não sendo possível sua existência sem que haja pelo menos um anel aromático principal. As seis posições seguintes, de N6 até N1, denotam estruturas de anéis naftênicos. Enquanto N6 e N5 denotam hidrocarbonetos cíclicos principais de 6 e 5 carbonos respectivamente, as posições de N4 até N1 denotam anéis naftênicos que se ligam a outros anéis, e como no caso das posições A4 e A2 não podem existir sem que haja um anel principal, neste caso naftênico ou aromático.

A décima posição, R, denota uma cadeia alifática acíclica que pode ser independente ou ser uma ramificação das estruturas anteriores. A posição br denota quantos carbonos são ramificações do tipo metil dentro dessa estrutura e a posição me denota quantos carbonos são ramificações do tipo metil ligados aos anéis aromáticos e ou naftênicos. A posição IH denota insaturações nas posições R e N6 até N1, e a posição AA denota ligação entre anéis aromáticos e ou naftênicos na forma de ponte, como por exemplo no caso fenilbenzeno. As demais posições denotam estruturas de enxofre, oxigênio e nitrogênio presentes seja na forma de heteroátomos ou na forma de ramificações, presentes nas estruturas anteriores. A Figura 1 representa alguns exemplos de estruturas moleculares com seus respectivos vetores.



A6	A4	A2
1	2	1

N6	N1	AA	NN
2	1	1	1

A6	R	br	me
1	7	1	1

Figura 1: Moléculas e seu vetor SOL, as posições omitidas possuem valor nulo.

Notação SMILES

A notação SMILES (*Simplified molecular-input line-entry system*) é um tipo de notação linear que descreve compostos químicos através de caracteres na codificação ASCII, podendo representar estruturas presentes na maioria das moléculas orgânicas como estruturas cíclicas, lineares, ramificações e até mesmo aromaticidade. O sistema se beneficia da notação linear para representar as moléculas em forma de grafos onde cada nó representa um átomo e cada caminho representa uma ligação[2].

Os átomos usualmente presentes em moléculas orgânicas são representados por suas abreviações químicas, outros átomos, como metais, por exemplo, também são representados dessa forma, entretanto devem ser colocados entre colchetes e acompanhados por suas cargas quando não nulas. Outra característica interessante desse sistema é que os átomos de hidrogênio ficam implícitos em moléculas orgânicas, respeitando a regra do octeto, não sendo necessária a escrita dos átomos de hidrogênio.

As ligações são representadas por diferentes símbolos, sendo os principais deles: o traço simples, -, que representa ligações simples e usualmente é implícita, não sendo necessária a sua escrita; a igualdade, =, que representa ligações duplas; a cerquilha, #, que representa ligações triplas; dois pontos, :, que representam ligações aromáticas. Devido ao caráter linear, as ramificações de uma estrutura principal devem ser colocadas entre parênteses, precedidas do tipo de ligação, e antecedidas pelo átomo a qual elas estão ligadas.

Estruturas cíclicas são representadas por trechos de texto que contenham os átomos presentes ordenadamente, onde o primeiro e o último átomos devem vir enumerados pelo mesmo número que representa que aquela estrutura é fechada. Anéis aromáticos são representados como estruturas cíclicas, onde os caracteres de átomos podem vir intercalados por ligações simples (traços) e duplas (igualdades) ou pelo símbolo dois pontos, outro modo de representação é escrever os caracteres do anel em sua grafia minúscula. Na Figura 2 são encontrados exemplos de moléculas com a respectiva representação SMILES.

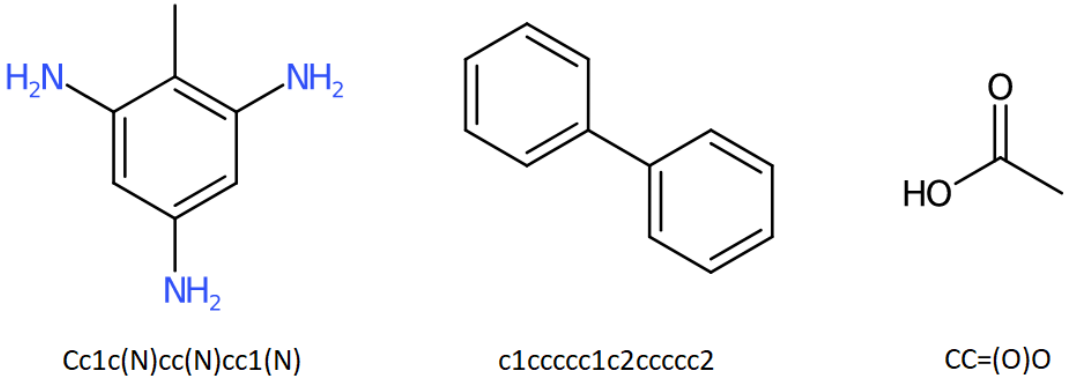


Figura 2: Moléculas e sua notação SMILES.

Metodologia

No método SOL as moléculas podem ser reconstruídas por incrementação, partindo das estruturas principais, anéis aromáticos, naftênicos e cadeias alifáticas, para as outras estruturas. Assim convém desenvolver algoritmos que traduzam essas estruturas para a notação SMILES e fornecer essa notação para um programa que permita fazer a representação gráfica da molécula. A linguagem escolhida para programação do algoritmo foi Python devido à sua facilidade de compreensão e escrita, à existência de bibliotecas de uso livre que representam moléculas e devido à sua ampla comunidade.

Analisando algumas moléculas e sua notação SMILES foi possível, através de testes em um representador de moléculas pré-existente compatível com a notação, identificar padrões na construção das estruturas principais. Esses padrões foram a base para a elaboração das funções responsáveis por cada estrutura presente no vetor SOL. As funções responsáveis por cada estrutura recebem o vetor SOL como entrada, com o número de sua estrutura correspondente bem como o número de estruturas das quais elas dependem, e, a partir dessas entradas, reproduzem os padrões identificados anteriormente. A partir dessas funções individuais, foi construída uma função principal que percorre o vetor sol e constrói um segmento de texto correspondente por incrementações e modificações.

Inicialmente, a construção da notação em SMILES foi feita partindo diretamente do vetor SOL para o segmento de texto correspondente. Entretanto, essa se mostrou uma solução inviável tendo em vista que outras estruturas do vetor SOL, como as de substituição por exemplo, só poderiam ser alocadas em locais específicos. Outro caminho para construção foi tomado, onde as moléculas eram representadas por listas de átomos que, por usa vez, eram representados por dicionários, os quais são estruturas de dados semelhantes às listas, porém seus índices não são ordenados e podem ser de diferente tipos de dados.

Cada átomo possui em seu dicionário as seguintes chaves:

- o símbolo atômico, que indica qual átomo o dicionário representa;
- prefixo e sufixo, que indicam se o átomo é o primeiro ou último de uma ramificação, portanto vindo acompanhado de parênteses;
- ligações, que indicam a ligação com o átomo anterior;
- enumeração, que indica se o átomo é o primeiro ou o último de uma estrutura fechada;
- liberdade, que indica se o átomo pode ou não receber ramificações.

Essa organização em forma de dicionários tornou mais simples a identificação de quais estruturas da base da molécula poderiam receber outras estruturas. Na Figura 3 tem-se o fluxograma que ilustra de forma simplificada o processo de construção das moléculas partindo do vetor SOL até a notação SMILES.

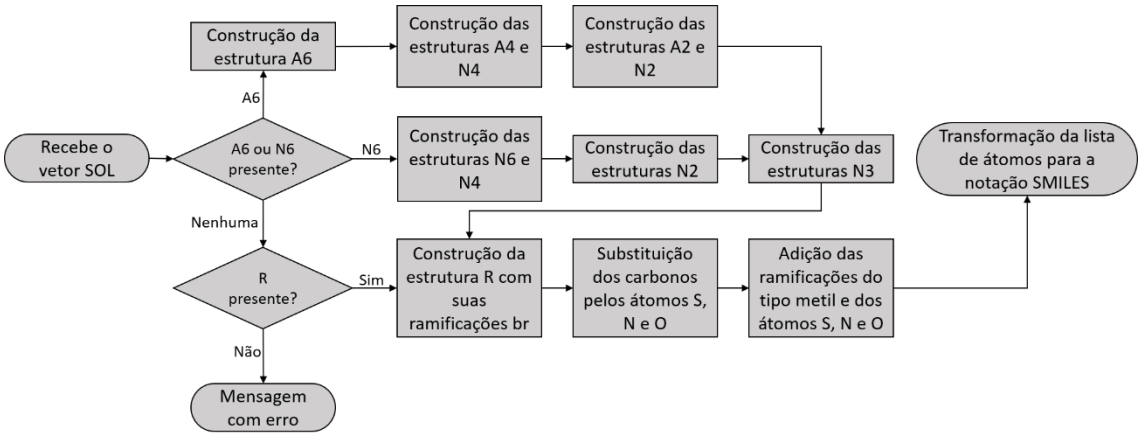


Figura 3: Fluxograma da tradução de um vetor SOL para notação SMILES.

A partir da notação SMILES é possível representar as moléculas em sua fórmula esquelética através de bibliotecas e pacotes disponíveis para a linguagem Python, um deles, o RDKit, principalmente por ser uma biblioteca de código aberto e de fácil uso.

Resultados e perspectivas

Com o algoritmo apresentado foi possível representar com sucesso moléculas simples. Além disso também foi possível gerar várias representações de forma rápida, com centenas de imagens na extensão .PNG em sengo geradas em poucos minutos, ou mesmo na extensão .SVG, que descreve imagens através de vetores gráficos. Na figura 4, conseguimos ver algumas moléculas representadas pelo algoritmo.

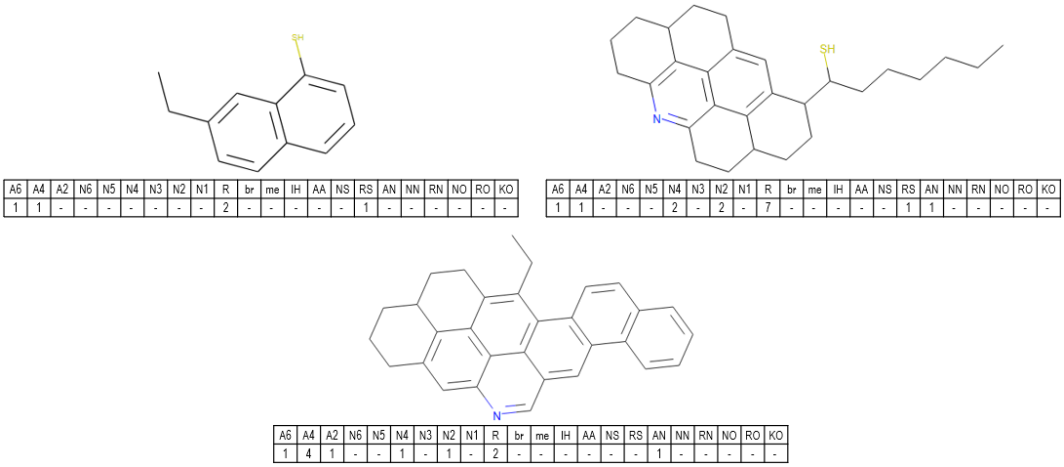


Figura 4: Representações moleculares geradas pelo algoritmo e seus respectivos vetores SOL.

Moléculas mais complexas com estruturas do tipo N1 e AA do vetor SOL encontram-se em desenvolvimento, principalmente pelo fato dessas estruturas estarem presentes quando há mais de um anel principal, isto é, mais de uma estrutura A6, N6 e/ou N5. Quando isso ocorre, é necessário que as outras estruturas A4, A2 e N4 até N2 sejam alocadas aleatoriamente entre as estruturas principais presentes, A6, N6 ou N5. As funções que tratam das estruturas N1 e AA estão sendo desenvolvidas com base no fluxograma, onde cada estrutura A6, N6 ou N5 receberá aleatoriamente as outras estruturas do vetor SOL e serão ligadas entre si pela estrutura AA e pela estrutura N1 quando presente.

Outra limitação atual do algoritmo são moléculas de alta complexidade, chamadas de moléculas multinucleares, onde várias moléculas simples, chamadas de núcleo, são ligadas entre si através de cadeias alifáticas. Essas moléculas apresentam um problema de sobreposição, impedindo a sua representação pela característica bidimensional da fórmula esquelética, fazendo que algumas partes das moléculas sejam sobrepostas durante a representação gráfica. A solução proposta até o momento foi a representação individual das moléculas simples, acompanhadas de uma indicação destacada dos átomos que estão ligados aos outros núcleos.

O algoritmo em desenvolvimento mostrou-se eficiente na representação de um número grande de moléculas simples. Completadas as funções que representam estruturas mais complexas, será possível representar moléculas multinucleares. Outra funcionalidade em desenvolvimento é a elaboração de uma interface simples de acesso a arquivos contendo os vetores SOL e que permita entrada do usuário. Após o desenvolvimento das principais funcionalidades do programa, serão testadas as funções e a interface em busca de melhorias de seu funcionamento, facilitando a instalação e o uso do programa para os usuários.

Agradecimentos

A Faculdade de Engrenharia Química e ao Prof. Dr. Dirceu Noriler pela oportunidade oferecida e orientação das diretrizes do projeto. Aos membros do Laboratório de Pesquisa em Processos Químicos e Gestão Empresarial (PQGe), em especial aos membros Tarcísio Dantas e Karina Klock pelo auxílio no decorrer e acompanhamento do projeto. Ao companheiro de projeto Lucas Henrique pela ajuda na integração do algoritmo e desenvolvimento da interface. Aos engenheiros da Petrobrás pela ajuda no entendimento do método SOL. À Petrobrás pela concessão de bolsa através da FUNCAMP durante o desenvolvimento do projeto.

Referência Bibliográfica

[1] J. Chen, Z. Fang, T. Qiu. **Molecular reconstruction model based on structure oriented lumping and group contribution methods**. Chinese Journal of Chemical Engineering. Beijing, China. Setembro de 2017.

[2] Daylight Chemical Information Systems, Inc. **Daylight Theory Manual**. Califórnia, EUA 2011. Disponível em: https://www.daylight.com/dayhtml/doc/theory/. Acesso em: maio de 2020.

[3] R.J. Quann, S.B. Jaffe, **Structure-oriented lumping: describing the chemistry of complex hydrocarbon mixtures**, Ind. Eng. Chem. Res. 31 (11) (1992) 2483–2497.

[4] RDKit-discuss - Mailing list for discussion, questions and answers. Disponível em: https://sourceforge.net/p/rdkit/mailman/rdkit-discuss/. Acesso em: junho de 2020.