Predição de fragmentação de espectros de massa utilizando energia e descritores de ligações.

Marcus Tullius Scotti (PQ)[1], Luciana Scotti (PQ)[2*], João Aires-de-Sousa (PQ)[3] *mtscotti@gmail.com

1 – Universidade Federal da Paraíba, Departamento de Engenharia e Meio Ambiente, Campus IV;
2 – Universidade Federal da Paraíba, Centro de Ciências de Saúde, Campus I;
3 – Universidade Nova de Lisboa, Departamento de Química, Faculdade de Ciências e Tecnologia, Portugal;

Palavras Chave: *Descritores de ligações, Random Forest.*

Introdução

A espectrometria de massas (EM) é uma técnica analítica para a investigação de moléculas ou misturas complexas que tem se desenvolvido drasticamente nos últimos anos, sendo cada vez mais empregada na área de produtos naturais. No âmbito da metabolômica de plantas, EM acoplada a técnicas de separação, tais como cromatografia líquida de alta eficiência ou cromatografia gasosa têm sido amplamente exploradas. Atualmente, o mercado global de EM representa uma faixa de R$ 7 bilhões. Este trabalho tem como objetivo desenvolver métodos *in silico* para auxiliar a determinação estrutural a partir de dados de EM.

Resultados e Discussão

Inicialmente foi coletado do banco de dados Massbank 5450 spectros de massa obtidos a partir de diversos métodos. Foram geradas a partir dos códigos SMILES as respectivas estruturas 2D (duas dimensões) em formato smi, posteriormente foram convertidas para o formato .sdf, utilizando o módulo Standadizer do programa JChem 6, (http://www.chemaxon.com). Para a automação de todo o procedimento e obtenção dos resultados foram criados e utilizados scripts em linguagem Python 2.7 e "workflows" com o software Knime 2.7.1. Para a geração de fragmentos foi utilizado o programa Met-Frag. Para indicar quais ligações são quebradas. Inicialmente, as estruturas das moléculas são fragmentadas sistematicamente e estes são comparados com os gerados pelo programa Met-Frag. Verifica-se dessa forma qual a ligação ou ligações quebradas responsáveis pelos fragmentos. Para a geração dos descritores de energia foi utilizado o programa computacional BDE v 2.5.7 que prediz as energias de ligações. Este software utiliza modelos, Random Forest (RF) ou Redes Neurais que foram treinados previamente utilizando as energias de ligações que foram previamente calculadas por DFT para milhares de estruturas.[1] Foram também utilizados os descritores de ligações desenvolvidos pelo mesmo grupo de pesquisa.[1] Estes descritores são baseados em esferas em torno das ligações. A partir destas camadas são contabilizados os tipos de átomos e ligações presentes dentro de cada esfera. Para a série de treino e teste foram selecionados aleatoriamente 4550 e 1000 espectros respectivamente para treino e teste. Após a geração dos modelos RF, as taxas de acerto para as ligações que quebram e não quebram para a série de treino, validação cruzada e teste, foram acima de 84,3 % para as que quebram e 96,7 % para as ligações que não quebram (tabela 1), portanto, os resultados do modelo mostram que a especificidade, verdadeiro negativo, é maior que a sensibilidade (verdadeiro positivo). A acurácia (predição geral) para a série de teste foi de 94,9%.

Tabela 1. Dados de acerto para as séries de treino, validação cruzada (utilizando 10 grupos) e teste.

	Treino		Crosss	Teste	
	Número Ligações	Acerto %	Acerto %	Número Ligações	Acerto %
N Q	158233	97,5	96,7	34989	96,7
Q	25850	90,7	85,2	5776	84,3
T	184083	96,6	95,1	40765	94,9

NQ – não quebra; Q – quebra; T – Total.

Das 1000 moléculas utilizadas na série de teste, o modelo conseguiu predizer em 371 todas as ligações que quebravam e que não quebravam e 782 acima de 90% das ligações. Se o limite for menos restritivo (80%) o número de compostos sobe para 918.

Conclusões

A abordagem desenvolvida utilizando as energias de ligações e descritores de ligação utilizando RF está predizendo significativamente as quebras de ligações por espectrometria de massa. Este método pode ser utilizado como ferramenta de predição de padrões de fragmentação ou incrementar a acurácia dos já disponíveis e utilizados atualmente, minimizando custos relativos ao tempo e consequentemente auxiliando a determinação estrutural de compostos.

Agradecimentos

Ao CNPq pelo auxílio financeiro.

[1] Qu X.; Latino, D. ARS.; Aires-de-Sousa, J. *J. Cheminfor.* **2013**, 5:34.