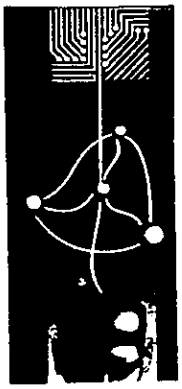

1° Congresso Brasileiro de Redes Neurais

Escola Federal de Engenharia de Itajubá
Itajubá, 24 a 27 de outubro de 1994

REDES NEURAIS NA QUÍMICA

Aguinaldo Robinson de Souza, Adriana Maria Francisco (Departamento de Química), Leandro Vernaschi de Mello, Rodrigo Alves Marson, Natália Moniwa (Departamento de Computação), Universidade Estadual Paulista, Bauru, S.P., 17030-360.

Resumo

Neste artigo são apresentadas algumas aplicações recentes de modelos de redes neurais na química. A enfase é dada ao algorítmo de "back-propagation" e mapas de Kohonen. As áreas da química que são abordadas são: Química Orgânica, Química Analítica e Química Biológica. São apresentados estudos sobre relações entre a estrutura química de um molécula e seu espectro de infravermelho; predições de reatividade de ligações químicas; estudos sobre a previsão da estrutura secundária de proteínas e estudos sobre relações quantitativas estrutura - atividade biológica.

I - Introdução.

Na química, como em todas as outras ciências naturais, muitos estudos estão sendo feitos para o desenvolvimento de novos métodos, que possam melhorar o tratamento de dados obtidos experimentalmente. A abordagem dos estudos de redes neurais à este problema é básicamente um método que possa tratar sistemas de multivariáveis e multirespostas.

O interesse dos químicos, e também dos bioquímicos, engenheiros químicos e farmacêuticos, no estudo de redes neurais tem crescido rápidamente desde 1986. Este interesse foi despertado em resposta aos trabalhos de Rumelhart e colaboradores (1) e ao aparecimento de um artigo de revisão de Lippmann (2). Os trabalhos referenciados no presente artigo cobrem uma gama de aplicações bastante ampla. O denominador comum destes estudos é a elaboração de modelos. Quando a abordagem de rede neural é utilizada na química, o que geralmente é procurado é algum modelo que possa representar a transformação de um certo tipo de entradas (input) em um conjunto de saídas (outputs). A vantagem de uma rede neural é que esta pode ser utilizada para generalizações (dentro de certos limites).

O modelo geralmente é construído a partir de um banco de dados, por exemplo, um banco de dados em espectroscopia ou a sequência de amino ácidos em proteínas. Este banco de dados deve conter um certo número de correlações, e outras informações que não podem ser deduzidas de uma maneira direta, seja por cálculos teóricos ou métodos numéricos.

II - Correlação estrutura-espectro de infravermelho.

A elucidação da estrututa de compostos orgânicos está bastante relacionada ao métodos espectroscópicos de análise. No entanto, as relações entre a estrutura de uma molécula orgânica e os dados espectrais são geralmente muito complexos para serem expressos em equações explícitas.

Até o momento estas relações foram estudadas utilizando-se regras empíricas. A procura de métodos que implementem o desenvolvimento da relação estrutura-espectro de infravermelho pode, afordunadamente, ser conduzida para o estudo de bancos de dados, de resultados experimentais, computadorizados. Hoje em dia com os resultados experimentais bastante precisos, estes bancos de dados possuem informações muito importantes sobre a estrutura de moléculas orgânicas. Desta maneira, a relação entre estes resultados experimentais e a elucidação da estrutura de moléculas é uma das áreas bastante promissora para a aplicação das técnicas de redes neurais.

Estas relações estrutura-espectro, não está restrita sómente à espectroscopia de infravermelho, mas pode envolver outras técnicas, como por exemplo espectroscopia de massas ou espectroscopia de ressonância magnética nuclear de C^{13} (3,4).

Como um exemplo desta aplicação apresentamos o trabalho de Munk e colaboradores sobre a interpretação de espectros de infravermelho (5). Neste trabalho um intervalo do espectro de 4000 - 400 cm^{-1} foi dividido em 640 intervalos de comprimento 5.6 cm^{-1}. O valor da intensidade de transmissão de um dado intervalo foi dado pela equação:

$$x_i = 1.00 - (\%t) / 100.0 \qquad (1)$$

onde %t = % de transmissão.

Desta maneira a rede neural necessitou de 640 unidades de entrada. No entanto como este valor era muito grande e causou alguns resultados espúrios, este número foi reduzido para 256. Ao mesmo tempo o comprimento dos intervalos foi ajustado de tal maneira que fossem pequenos à baixas baixas frequências e grandes à altas frequências. A expressão que leva em conta a dependência do comprimento do intervalo i com a frequência, é dada pela equação abaixo:

$$i = 6.0\ (\text{frequência})^{0.5} - 120.0 \qquad (2)$$

Esta equação atribui um intervalo de frequência de 10 cm^{-1} (de 400 - 410 cm^{-1}) à unidade de entrada 1. No outro lado do espectro, esta equação atribui um intervalo de frequência de 20 cm^{-1} (de 3928 - 3948 cm^{-1}) à unidade de entrada 256. Estes intervalos de frequência são então distribuidos entre os picos do espectro; se um pico é encontrado, a sua frequência (o valor deve estar entre 0.000 e 1.000) é a entrada desta unidade, senão a entrada desta unidade é igual a zero.

A estrutura do composto é descrita em termos de 36 grupos funcionais: alcool primário, fenol, amina terciária, éster, etc., cada um representado por uma unidade de saída. Desta maneira o vetor a ser encontrado é um vetor binário com 36 variáveis, onde cada número 1 indica a presença do grupo funcional associado, e o número zero indica a sua ausência. Em geral, uma estrutura pode ter vários grupos funcionais, e desta maneira várias unidades de saída podem estar ativas simultâneamente. Após o treino de 14 diferentes redes, variando de 10 até 60 neurônios, 34 foi encontrado como o melhor número.

Devido à possibilidade de uma rápida associação nas redes neurais e também da habilidade em armazenar e tornar acessível grandes quantidades de dados em espaços multidimensionais, Zupan (6) especulou sobre a possibilidade da criação de um espectrômetro inteligente. A sua premissa foi que a quantidade de informações processadas por um espectrômetro de infravermelho, na sua vida média, não excede 1.6 Gbyte (100 dias,

vezes 200 espectros por dis, vezes 4000 valores por espectro, vezes 2 bytes para cada valor). Tal valor está dentro das capacidades de software e hardware existentes hoje em dia.

III - Estrutura secundária de proteínas.

Os polipepitídeos e as proteínas são constituídos de unidades elementares chamados aminoácidos Uma proteína é constituida, salvo alguns casos especiais, de sómente 20 aminoácidos.

Estes aminoácidos estão arranjados sequêncialmente numa proteina; a sequência exata é chamada estrutura primária. Esta sequênica linear enovela-se (folds) em uma única estrutura tri-dimensional, que contém as caracteristicas globais do que é chamada estrutura secundária. Existem três tipos de estruturas secundárias: hélice-α, estrutura-β, e novelo aleatório.

A estrutura secundária de uma proteína é de grande importância para a sua atividade biológica.

Desta maneira, existe muito interesse na predição da estrutura secundária de proteínas a partir do conhecimento de sua estrutura primária. O método mais utilizado é o de Chou e Fasman (7,8), que permite predizer, a partir da sequência de aminoácidos, se um dado aminoácido faz parte de uma α-hélice, uma estrutura-β, ou um novelo aleatório.

Nos últimos anos, muitos estudos foram feitos sobre a utilização de redes neurais para a predição de estruturas secundárias de polipepitídeos a partir da sequência de aminoácidos; os pioneiros neste campo foram Qian e Sejnowski (9).

A suposição básica dos trabalhos de Chou e Fasman e também de Qian e Sejnowski, é de que a identidade de um aminoácido e seus vizinhos determina a estrutura secundária de seus vizinhos. Um certo tipo de "janela" sobre um segmento de um polipepitídeo poderia em princípio fornecer detalhes sobre a estrutura secundária da cadeia inteira.

Na abordagem de redes neurais são necessários três neurônios de saída, um para cada estrutura: hélice-α, estrutura-β, e novelo aleatório. Qian e Sejnowski tentaram diferentes números de neurônios (0 - 80) e decidiram que o número ideal foi 40. O teste de treinamento consistiu de 106 proteínas, tendo um total de 18.105 aminoácidos. Cada um destes esteve acompanhado pela especificação do tipo de estrutura secundária inerente à este resíduo. Um dado aminoácido pode ser encontrado em diferentes tipos de estruturas em diferentes porteínas, ou mesmo em diferentes partes da mesma proteína.

Os trabalhos de Qian e Sejnowski despertaram grande interesse entre os pesquisadores que estudam a estrutura secundária de proteínas. A partir de então, foram publicados um número expressivo de trabalhos nesta área. No entanto, os resultados obtidos até o momento não são conclusivos, e muitos estudos estão sendo feitos neste campo, de interesse cada vez mais crescente entre os pesquisadores.

IV - Reatividade de ligações químicas.

A predição do mecanismo e dos produtos de uma reação química é um dos objetivos principais dos pesquisadores da área de Química Orgânica. Visto que as reações químicas são iniciadas pela quebra de uma ou mais ligações em uma molécula, o conhecimento da reatividade das ligações químicas, isto é das ligações mais fáceis de se quebrar, é indispensável para a predição de reações químicas.

As reações químicas são governadas por processos polares, onde o resultado é uma carga negativa sobre um átomo e uma carga positiva sobre outro átomo. Este processo é conhecido como heterólise.

Simon e colaboradores (10) treinaram uma rede neural para a predição da quebra de ligações em compostos alifáticos. Esta rede neural foi capaz de predizer com grande exatidão qual a ligação

que se quebraria mais fácilmente, e como as cargas iriam se distribuir entre os átomos da molécula.

Um banco de dados de 29 moléculas foi escolhido como um grupo representativo das variações estruturais do grupo de moléculas alifáticas; estas moléculas possuem 385 ligações capazes de 770 modos potenciais, de quebra heterolítica. Considerando sómente as ligações simples (por exemplo: C-H em um grupo metílico) temos 373 modos diferentes de quebra. Destes 373 modos, uma série de 149 modos foram selecionados e divididos em 43 reativos e 106 não reativos. A quebra de uma ligação química é influenciada por uma variedade de efeitos energéticos, eletrônicos ou estéricos, como por exemplo: distribuição de cargas, indução, ressonância, polarizabilidade, energia de dissociação de ligações, etc.

Dentre os fatores que governam uma reação química temos: a diferença na carga total, Δq_{tot}, a diferença na carga-π, Δq_{π}, a diferença na eletronegatividade-σ, $\Delta \chi_{\sigma}$, a medida da polaridade da ligação, Q_{σ}, a estabilização por ressonância, $R^{\pm}$, a polarizabilidade da ligação, α_b, e a energia de dissociação da ligação, BDE. Os valores destas variáveis são calculados e atribuídos às ligações usando um programa chamado PETRA (Parameter Estimation for the for Treatment of Reactivity Applications) (11).

O objetivo principal nestes estudos é relacionar estas variáveis á uma classificação de reatividade de uma ligação química. A classificação da reatividade de ligações químicas é claramente um problema de muitas variáveis, e portanto bastante adequado para ser estudado utilizando os modelos de redes neurais.

A primeira abordagem ao problema é classificar os modos de quebra como reativos ou não reativos através de uma rede neural de duas camadas, utilizando o aprendizado por "back-propagation". O número de unidades de entrada é definido como sendo igual a sete, o número de efeitos controladores da reatividade. Estes possuem diferentes valores, um deles possue valores no intervalo de -0.2 a +0.2, outro de 200 a 500; visto que as unidades de entrada esperados devem estar entre 0 e 1, cada unidade de entrada deve ser escalonada (scaled) entre os seus valores máximos e mínimos. A classificação das unidades de saída é: 0 para modos não reativos, e 1 para modos reativos.

O estudo deste sistema foi realizado utilizando uma rede de Kohonen (9x9), contendo 81 neurônios que podem ser considerados como "sub-espaços". Quando o banco de dados possue ligações com uma depência similar sobre as sete variáveis de controle, esta rede irá mapear então estas ligações sobre o mesmo neurônio.

A rede é estabilizada após 30 ciclos de treinamento, isto é, após os 373 modos de quebra de ligação terem sidos analisados na rede por 30 vezes. Seis neurônios estão vazios, 56 possuem modos classificados e 19 estão ocupados por modos não classificados. A rede de Kohonen utiliza uma técnica de aprendizado não supervizionada visto que não utiliza a informação da classe quando está no processo de aprendizado. Esta rede possui uma quantidade de informação, de interesse dos químicos, muito grande.

Encontrou-se que os modos reativos (**R**) formam um agregado no centro do mapa. Esta é uma indicação de uma auto-organização durante o processo de aprendizado da rede de Konohen, que percebe a similaridade entre certos tipos de modos e coloca-os então no mesmo neurônio. Esta rede também leva ao reconhecimento da similaridade de todos os modos reativos colocando-os em neurônios vizinhos, formando desta maneira agregados de neurônios com modos reativos.

V - Relações Quantitativas Estrutura - Atividade Biológica (QSAR).

O campo de estudo de relações quantitativas estrutura - atividade biológica

foi introduzido no início dos anos 60 com os trabalhos pioneiros de Hansch e colaboradores (12,13). Em uma sequência de publicações estes pesquisadores demonstraram que a atividade biológica de certos compostos químicos obedece a uma função matemática das suas características fisico-químicas, como por exemplo: hidrofobicidade, forma e propriedades eletrônicas. Estes métodos estão sendo largamente adotados pelas indústrias farmacêuticas e agroquímicas (14). A pesquisa de tais relações é uma das aplicações mais importantes das técnicas de modelagem molecular.

A correlação da estrutura química de fármacos com suas atividades biológicas é de particular interesse, devido principalmente ao elevado custo de novos fármacos. Uma predição quantitativa, confiável, de sua atividade antes do novo fármaco ser sintetizado é de grande interesse para os laboratórios farmacêuticos que realizam estas sinteses.

Podemos citar alguns trabalhos que mostram a relevância do tema. Temos o trabalho de Marzona (15) sobre complexos de inclusão de esteróides com ciclodextrinas. A inclusão de uma molécula de fármaco na ciclodextrina pode alterar consideravelmente suas características, principalmente no plano farmacotécnico (16).

Os estudos de Tetko e colaboradores (17), sobre a aplicação de redes neurais artificiais no estudo de derivados de carboquinonas, também mostram a crescente importância destes modelos no estudo de QSAR. Os derivados das carboquinonas são amplamentes utilizados como agentes anti-tumor, por exemplo contra a leucemia L-1210 (18).

Como um exemplo típico da aplicação de modelos de redes neurais em QSAR discutiremos o trabalho de Aoyama e colaboradores (19,20).

O banco de dados neste estudo envolveu modificações no esqueleto básico da carboquinona. Muitas carbiquinonas exibem graus variados de atividade anticarcinogênica. Este estudo de QSAR teve como objetivo predizer a dose mínima de fármaco necessária para produzir uma extensão de 40% na vida de cobaias, ratos que foram inoculados com células de leucemia L-210.

Esta dose efetiva mínima depende da concentração, C, da substância necessária para obter o efeito necessário, e é representada como log (1/C). Quanto mais efetiva for o fármaco menor será a concentração necessária.

Como era esperado encontrou-se que a atividade anticarcinogênica depende da identidade dos substituintes R^1 e R^2, na estrutura da carboquinona. Nas análises de regressão multilinear, correntemente usadas, estes substituintes são descritos por variáveis fisico-químicas que descrevem a influência combinada dos substituintes R^1 e R^2. A atribuição dos substituintes como R^1 e R^2 é baseada nas suas refratividades molares: $MR_1 \leq MR_2$. Neste estudo foram utilizados onze diferentes substituintes R^1 (consistindo básicamente de grupos alquila de cadeia curta como o grupo metila, etila e propila) e cerca de 30 diferentes substituintes R^2, estes de cadeia mais longa e carregando funcionalidades adicionais, como $-CH_2CH_2OCH_3$ e $-CH(OCH_3)CH_2OCONH_2$.

A rede consistiu de seis unidades de entrada e um neurônio de saída. Após várias tentativas 12 neurônios foram necessários na "hidden layer". Esta rede de (6 x 12 x 1) neurônios foi treinanda com 35 carboquinonas utilizando o algorítmo de "back-propagation". Os valores de log (1/C) foram então comparados com os resultados obtidos através de análise de regressão multilinear, com as mesmas 35 carboquinonas.

A atividade anticarcinogênica de 17 das carboquinonas foi predita com uma

maior precisão do que o estudo utilizando análise de regressão multilinear; para 6 compostos os resultados tiveram a mesma qualidade, e para 12 os resultados foram piores. De uma maneira geral os resultados utilizando redes neurais são significativamente (mas não dramáticamente) melhores do que a análise de regressão multilinear.

Aparentemente o problema estudado é bastante adequado para ser tratado por modelos lineares, mas a abordagem utilizando redes neurais pode ir um pouco mais além. Problemas de QSAR não lineares poderão ser melhores estudados quando modelados por redes neurais.

VI - Conclusões e perspectivas.

Uma das consequências das aplicações de redes neurais na química é de que somos forçados a reconsiderar o modo como representamos e interpretamos os dados obtidos experimentalmente. A representação dos dados é muito importante para a extração de informações. A abordagem de redes neurais, seja no aprendizado via mapas de Kohonen ou "back-propagation", colocou este fato em mais evidência ainda; a representação dos dados é crucial.

Acreditamos que a tendência nesta área de pesquisa será a de desenvolvimento de soluções dedicadas. Desta maneira os laboratórios industriais irão procurar patentear tais soluções, como já vem acontecendo na indústria farmacêutica, e a troca de informações sobre as pesquisas nesta área serão desta forma afetadas. Portanto, torna-se cada vez mais importante que os órgão governamentais, através de auxílios às Universidades, suportem as pesquisas nesta área.

VI - Referências Bibliográficas.

1) D. E. Rumelhart, G.E. Hinton, R.J. Willians, *Microstructures of Cognition, Vol 1, MIT Press*, Cambridge, (1988).
2) R. P. Lippmann, *IEEE ASSP Magazine*, April 4 (1987).
3) B. Curry, D.E. Rumelhart, *Tetrahedron Comput. Methodol.* **3** (1990) 213.
4) V. Kvasnicka, *J. Math. Chem.* 6 (1991) 63.
5) M.E. Munk, M.S. Madison, E.W. Robb, *Mikrochim. Acta [Wien]* II (1991) 505.
6) J. Zupan, *Anal. Chim. Acta*, **53** (1990) 53.
7) P.Y.Chou, G.D. Fasman, *Biochemistry* 13 (1974) 211.
8) P.Y. Chou, G.D. Fasman, *Biochemistry* 13 (1974) 222.
9) N. Qian, T.J. Sejnowski, *J. Mol. Biol.* 202 (1988) 865.
10) V. Simon, J. Gasteiger, J. Zupan, *J. Am. Chem. Soc.*, in press.
11) J. Gasteiger, M. Marsili, M.G.Hutchings, H. Saller, P. Löw, P. Röse, K. Rafeiner, *J. Chem. Inf. Comput. Sci.* **30** (1990) 467.
12) C. Hansch,R.M. Muir, T. Fujita, P. Maloney, E. Geiger, M. Streich, *J. Am. Chem. Soc.* **86** (1964) 2817.
13) C. Hansch, T. Fujita, *J. Am. Chem. Soc.* **86** (1964) 1616.
14) J.E. Ridings, D.T. Manallack, M.R. Sauders, J.A. Baldwin, D.J. Livingstone, SmithKline Beecham Pharmaceuticals, Welwyn, AL6 9AR, Herts, England, *Quant. Struct.-Act. Rel.* 12 (1993) 272.
15) M. Marzona, R. Carpignano, P. Quagliotto, *Quant. Struct.-Act. Rel.*12 (1993) 299.
16) A. Korolkovas, *Rev. Bras. Med.* 49 (1992) 509.
17) I.V. Tetko, A. I. Luik, G. I. Poda, *J. Med. Chem.* **36** (1993) 811.
18) M. Yoshimoto, H. Miyazawa, H. Nakao, K. Shinkai, M. Arakawa, *J. Med. Chem.* 22 (1979) 491.
19) T. Aoyama, Y. Suzuki, H. Ichikawa, *J. Med. Chem.* **33** (1990) 905.
20) T. Aoyama, Y. Suzuki, H. Ichikawa, *J. Med. Chem.* **33** (1990) 2583.