RESUMO:

O método recente de análise de cadeias peptídicas com o objectivo de determinar a sequência de aminoácidos, explora o carácter quase aleatório do fraccionamento de cadeia por agentes hidrolisantes. e a consequente sobreposição parcial dos segmentos obtidos.

A automatização do processo de reconstituição, desenvolvida no presente trabalho, constrói numa primeira fase um grafo orientado das relações observadas entre os segmentos e gera numa segunda fase, por percursos condicionados no grafo, possíveis sequências para a estrutura original.

O algoritmo desenvolvido foi orientado para funcionamento a nível de minicomputadores tendo sido testados casos realistas numa configuração reduzida.

No seguimento de trabalhos anteriores (Ref 1), têm sido desenvolvidos, no Departamento de Química da Universidade Nova de Lisboa, métodos de análise de peptídeos que levantam problemas interessantes para aplicação de métodos computacionais. Com efeito um peptídeo é uma cadeia constituída por unidades químicas elementares, os aminoácidos, ligados numa sequência bem determinada e característica de cada espécie química. o objectivo dos métodos analíticos é o de determinar completamente a referida sequência isto é, saber não só as proporções relativos dos aminoácidos e o seu peso total como numa formula química habitual, mas também em detalhe a ordem pela qual se sucedem os aminoácidos na cadeia.

O método experimental desenvolvido, e que se afigura, de bastante interesse para variadas classes de aplicações, consiste no ataque da cadeia peptídica por agentes hidrolizantes. Deste ataque resulta o fraccionamento quase aleatório da cadeia em segmentos menores. Tipicamente uma cadeia composta por uma sequência de 20 aminoácidos produzirá segmentos cujo comprimento é o de 2 a 5 aminoácidos. Desta forma, e com tratamento químico apropriado, produzem-se derivados susceptíveis de estudo directo por análise incremental dos seus índices de retenção em cromatologia gás-liquido e cuja análise se complementar por espectrometria de massa permite estabelecer a sequência de cada um dos segmentos menores produzidos.

Embora sujeito a ambiguidades provenientes de eventuais clivagens preferenciais, o método explora o carácter aleatório do fraccionamento e a consequente sobreposição parcial dos segmentos assim obtidos. A sequência original é então obtida procurando compatibilizar por sobreposição parcial e continuada todos os elementos obtidos na análise anterior.

O presente trabalho situa-se numa perspectiva de análise de cadeias lineares, isto a lacetes ou outras efíções estruturais mais complexas e procura automatizar o processo de reconstituição da sequência da cadeia original. Neste contexto pode representar-se cada aminoácido por um carácter num alfabeto básico, uma cadeia por uma sequência de caracteres "string", e utilizar as técnicas de manipulação de "string" em duas fases distintas da análise.

A primeira fase ocorre na altura da identificação dos dados fornecidos pela cromatografia gás-liquido. O processo analítico fornece para cada segmento menor um índice de retenção que é aditivo relativamente aos aminoácidos dos intervenientes. Torna-se pois necessário identificar as combinações de aminoácidos que fornecem índices de retenção dentro da margem de erro consentida pelos valores experimentais. A automatização deste processo é trivial enquanto os segmentos menores não contiverem mais do que cinco aminoácidos e se poder consentir na geração sucessiva de todas as possíveis combinações. Para números superiores, felizmente infrequentes, tem que se limitar o volume de geração de combinações no hiperespaço correspondente a fim de fornecer tempos razoáveis de computação.

Na segunda fase processa-se efectivamente a reconstituição da sequência original compatibilizando os segmentos menores observados, em termos de sobreposição parcial continuada. O algoritmo que a seguir se descreve elimina a necessidade de considerar todas as possíveis combinações, reduzindo

duma maneira drástica o tempo de cálculo para valores praticáveis em minicomputador.

O conjunto completo das subsequências obtidas pela análise cromatográfica e pela espectrometria de massa, é fornecido ao computador que começa por analisar as relações existentes entre os segmentos com o objectivo de determinar todas as possibilidades de sobreposição. A maneira mais simples de exprimir as relações de sobreposição consiste na construção dum grafo orientado em que cada nó representa um segmento. Sempre que um segundo na permite, por sobreposição continuar um primeiro. estabelece-se uma ligação orientada do primeiro para o segundo. A análise completa dos segmentos fornecidos gera por este processo uma estrutura complexa que materializa o referido grafo.

Seguindo as ligações estabelecidas obtêm-se sequências que são compatíveis com os elementos observados.

A exploração da estrutura criada consiste portanto na geração de percursos no grafo que restituem em sequências com propriedades relacionadas com as condições particulares do problema específico em estudo. Numa perspectiva de geração da cadeia completa interessam obviamente apenas as sequências que permitam enquadrar todos os segmentos fornecidos. Por outro lado as sequências geradas tem que ser compatíveis com outras propriedades químicas da cadeia em estudo e assim as sequências geradas são submetidas a testes de validade estrutural antes de aparecerem como alternativas possíveis para a estrutura do peptido original.

Dum ponto de vista pratico, porém, é preferível fornecer um processo interactivo em que algumas decisões são tomadas pelo próprio investigador que assim utiliza o computador como instrumento auxiliar na sua pesquisa de estrutura original. Tipicamente o investigador requerer; sequências cujo comprimento se encontre entre limites dados, a fim de examinar eventualmente mais do que uma estrutura possível.

Por outro lado este aspecto interactivo é necessário na destrinça de situações experimentais em que se verifique clivagem preferencial. Nesta situação não existe nenhum segmento que por sobreposição permita fazer a ligação nesse ponto. Este aspecto pode inclusivamente manifestar-se na própria estrutura do grafo permitindo o aparecimento de nós que não apontados doutros e até mesmo a existência de subgrafos. Muito embora o problema possa ser evitado permitindo a concatenação de segmentos, e não apenas a sua sobreposição parcial, tal estratégia não parece aconselhável em face do grande aumento que provoca nas possibilidades de combinação e do conseqüente aumento do tempo de cálculo. É pois aparentemente preferível deixar ao investigador o trabalho de guiar o processo de geração utilizando subconjuntos de dados quando estas situações anormais ocorrem.

o algoritmo desenvolvido foi testado em BASIC numa configuração com cerca de 20 kbytes de espaço de utilizador e com dados realistas provenientes da análise da cadeia 8 da Insulina (bovina).

Note-se que o algoritmo foi testado com uma parte coerente dos dados da análise da cadeia referida. Dada a complexidade do grafo inicial apenas se mostra na fig. 1, a titulo de exemplo o grafo correspondente e um subconjunto dos dados utilizados.

O Quadro 1 mostra a notação utilizada os dados fornecidos e as sequências obtidas com comprimento requerido entre 15 e 20 que usam todos os segmentos dados. Os resultados mostram no caso das sequências de comprimento 20 a ocorrência de soluções indescerníveis apenas com os dados da análise química normal visto que apenas envolvem alteração da posição de grupos de aminoácidos sem alteração portanto das suas proporções relativas. De notar que a sequência' mais curta (comprimento 16) é a que corresponde a estrutura conhecida da cadeia o da insulina (bovina).

A implementação eficiente do algoritmo requer uma liberdade apreciável no encadeamento de chamadas

de subrotina ou na imbricação de ciclos de computação apontando para a recursividade.

Em regra estes requisitos não são satisfeitos pelas linguagens tradicionais dos minicomputadores. Na ausência de linguagens estruturadas como o Algor e quando FORTRAN ou O BASIC existentes sofram de limitações no nível de profundidade de encadeamento ha que programar explicitamente a referida Imbricação. Porém o conseqüente aumento da utilização de espaço em memória pode ser minorado fazendo funcionar o algoritmo em dois ou mais passos produzindo seqüências de dimensão intermédia que no passo seguinte se submetem como dados. A escolha dependerá das circunstancias particulares e tem que ter em conta a comparação entre o espaço adicional de memória requerido pelo aumento do programa e o espaço adicional necessário para armazenar o grafo intermedio, agora de dimensão muito superior à do grafo inicial.

O problema foi originado pelo Professor Higinaldo J.C. das Neves do Departamento de Química da FCT, UNL a quem se agradece a disponibilidade sempre encontrada e as estimulantes e esclarecedoras discussões mantidas ao longo da evolução do trabalho, assim como a utilização dos dados da insulina acima apresentados.