III Simpósio Doutoral
do Departamento de Informática
da Universidade do Minho

SDDI 2006

Fevereiro de 2006

"Extracção de Padrões em dados Bioquímicos"

Doutorando: Pedro Gabriel Dias Ferreira
pedrogabriel@di.uminho.pt

Orientador: Paulo Jorge Azevedo
pja@di.uminho.pt

Secção A - Identificação

Doutorando
Pedro Gabriel Dias Ferreira (pedrogabriel@di.uminho.pt)

Título de Tese
"Extracção de Padrões em dados Bioquímicos"

Orientador
Paulo Jorge Azevedo
Universidade do Minho
Dep. de Informática
pja@di.uminho.pt

Data Início
Oficial: Prevista: Outubro de 2003

Data Final
Oficial: Prevista: Março de 2007

Secção B - Resumo

Resumo
Com o desenvolvimento da Biologia Molecular surgiu um crescimento exponencial de informação biológica. Este crescimento de dados biológicos apresenta-se em dimensões de complexidade e volume. O projecto do Genoma Humano é um exemplo de um enorme contributo para o crescimento destes dados, onde estão representados 3 biliões de bases de DNA que constituem o genoma humano. Como consequência, é já possível o acesso às sequências de cerca de 100,000 proteínas. Assim, podemos considerar que estamos na era pós-genoma onde, após algumas décadas de esforço na recolha e consolidação de dados genéticos, é agora a altura de passar a um outro passo de grande importância que é a análise destes dados. A esperança é de termos acesso à tecnologia que permita aprender mais sobre comportamento genético (gene expression), interacção entre proteínas e outros mecanismos biológicos. Será fácil perceber que analisar esta quantidade megalómana de dados é um enorme desafio. Não só pela sua complexidade e pelos múltiplos factores correlativos mas também pela constante evolução do entendimento dos mecanismos biológicos. As aproximações clássicas da análise de dados biológicos já não são suficientes (nem eficientes) para o tipo de informação que se pretendem extrair, nomeadamente quando tomamos em conta os numerosos e complexos mecanismos biológicos que se pretende estudar. Na verdade, estes métodos clássicos fazem uso de um pequeno número de parâmetros para representar os vários factores de correlação envolvidos nos mecanismos biológicos em estudo. Surge assim a necessidade de desenvolver ferramentas computacionais e aproximações de alto desempenho para apoio à análise de dados biológicos. Estas irão certamente contribuir para uma melhor compreensão das ``ligações/correlações" que existem entre as estruturas e os padrões funcionais nas sequências biológicas (DNA, RNA e proteínas) e também entre os mecanismos genéticos e os bioquímicos. As técnicas de Extracção de Conhecimento (KDD - Knowledge Discovery in Databases), na qual estão inseridas as técnicas de Mineração de Dados e de Bioinformática, tentam responder a estes desafios. Estas técnicas funcionarão como um passo de pré-processamento na análise de dados

biológicos. As relações e padrões extraídos nesta fase, serão posteriormente usados na verificação de hipóteses ou na previsão e explicação do conhecimento adquirido. Um caso particular de Mineração de Dados, é o da extracção de padrões sequenciais (motifs) em conjuntos de sequências (proteínas ou ADN) funcionalmente relacionadas. Os motifs são elementos temáticos recorrentes encontrados em várias moléculas. A partilha de um dado motif por várias moléculas será indicador de um eventual antepassado comum, podendo corresponder a segmentos de resíduos (ácidos nucleicos ou amino-ácidos) conservados durante a evolução devido a importantes tarefas funcionais ou estruturais das moléculas. O espectro de aplicação destes padrões sequenciais vai desde a classificação, agregação (clustering), anotação de famílias de sequências até a um suporte na melhor compreensão da relação entre sequência, estrutura e função. Dado as características das sequências biológicas e a natureza combinatória da extracção automática de padrões sequenciais, os algoritmos que atacam este problema resultam computacionalmente pesados. Numa primeira fase desta tese propomos o estudo e o desenvolvimento de algoritmos para a extracção de padrões sequenciais frequentes, i.e. padrões que ocorrem num número de sequências maior ou igual ao definido pelo utilizador. Numa segunda fase propomos o estudo do impacto e das aplicações dos padrões obtidos a partir dos algoritmos anteriormente descritos. Em concreto estudamos a aplicação dos padrões à tarefa de classificação de sequências de proteínas e o impacto de diferentes medidas de relevância estatística para determinação de padrões com significado biológico.

Objectivos Estratégicos

1. Estudo e Análise de algoritmos de mineração de padrões sequenciais;

2. Análise comparativa dos algoritmos provenientes da comunidade de Data Mining e da comunidade Bioinformática, tendo em conta a sua adequação aos dados biomoleculares (sequências de proteínas/dna);

3. Desenvolvimento e optimização de um algoritmo de mineração de padrões sequenciais em base de dados de proteínas e que suporte as várias restrições e propriedades necessárias a este tipo de mineração;

4. Estudo da relação entre a relevância estatística e biológica dos padrões sequenciais;

5. Estudo de casos de aplicação dos padrões sequenciais, nomeadamente em tarefas de classificação de sequências;

6. Aplicação mediante uma eventual transformação dos dados, dos algoritmos de mineração a dados de séries temporais.

7. Estudo da aplicação e da combinação dos algoritmos de mineração de padrões sequenciais com outros algoritmos, nomeadamente de associação, para a extracção de padrões de natureza temporal;

Secção C – Contribuições

Principais contribuições técnico-científicas

Proposta de um algoritmo, denominado gIL, para a mineração de padrões sequenciais em bases de dados de sequências de proteínas. O algoritmo resulta da combinação de algumas das técnicas mais eficientes utilizadas em mineração de sequências e de associações. O algoritmo exibe uma boa adaptabilidade, permitindo a incorporação directa de diferentes propriedades no processo de mineração, nomeadamente a extracção de padrões rígidos e flexíveis e a introdução de restrições. A avaliação experimental demostrou um eficiência superior aos algoritmos mais avançados nesta tarefa. Espera-se que este algoritmo venha a ser utilizado por bioquímicos nos seus processos de análise de sequências de proteínas. De notar que tal algoritmo pode ser utilizado em outros domínios, como por exemplo na mineração de padrões de utilizações da web.

Proposta de um algoritmo de classificação de sequências de proteínas. Este algoritmo utiliza os padrões rígidos que se podem extrair das sequências a classificar, a partir do algoritmo descrito no parágrafo anterior. A relevância de cada padrão é combinada através de um classificador Bayesiano, permitindo obter um valor global de similaridade da sequência a classificar em relação a cada família de sequências presente no conjunto de dados. Este classificador é multi-classe e não necessita de transformação dos dados e de conhecimento adicional. A avaliação experimental deste classificador demonstra uma eficiência equivalente aos métodos que definem o "estado-da-arte" nesta tarefa. Espera-se que o algoritmo possa vir a ser utilizado em tarefas específicas de classificação de proteínas, nomeadamente em casos em que os restantes algoritmos demonstrem uma qualidade de classificação inferior.

Proposta de um método que combina algoritmos de mineração de sequências com a mineração de associação de forma a extrair padrões com carácter inter-dimensional. Assim, quando aplicado a dados com dimensões espaciais e temporais, esta combinação de técnicas irá permitir a extracção de padrões *espacio-temporais*. Espera-se que este método possa vir a ser útil na análise de um largo número de domínios, como por exemplo na análise de dados das bolsas de valores, de base de dados de meteorologia ou em dados resultantes das estratégias de preços em operações no mercado retalhista.

Escrita de um tutorial que reúne os principais conceitos da biologia molecular e que poderá servir como introdução sucinta e objectiva para os cientistas da computação que desejem iniciar-se nesta temática.

Escrita de um tutorial que serve de introdução à temática da mineração de padrões sequenciais. Este tutorial começa por apresentar os conceitos básicos desta área de investigação, passando por uma descrição de alguns dos algoritmos e consequente comparação entre os diferentes algoritmos. Este documento poderá ser bastante útil para aqueles que desejem iniciar-se neste tema.

Publicações

Bioinformatics

2005

Pedro Gabriel Ferreira, Paulo Azevedo, ***"Protein Sequence Pattern Mining with Constraints"***, in Proceedings of 9th European Conference on Principles and practice of Knowledge Discovery in Databases, LNAI Springer-Verlag, (Porto, Portugal, October 2005).

Considering the characteristics of biological sequence databases, which typically have a small alphabet, a very long length and a relative small size (several hundreds of sequences), we propose a new sequence mining algorithm (IL). gIL was developed for linear sequence pattern mining and results from the combination of some of the most efficient techniques used in sequence and itemset mining. The algorithm exhibits a high adaptability, yielding a smooth and direct introduction of various types of features into the mining process, namely the extraction of rigid and arbitrary gap patterns. Both breadth or a depth first traversal are possible. The experimental evaluation, in synthetic and real life protein databases, has shown that our algorithm has superior performance to state-of-the art algorithms. The use of constraints has also proved to be a very useful tool to specify user interesting patterns.

Pedro Gabriel Ferreira, Paulo Azevedo, ***"Protein Sequence Classification Through Relevant Sequence Mining and Bayes Classifiers"***, in Proceedings of 12th Portuguese Conference on Artificial Intelligence, LNAI Springer-Verlag, (Covilhã, Portugal, December 2005).

We tackle the problem of sequence classification using relevant subsequences found in a dataset of protein labelled sequences. A subsequence is relevant if it is frequent and has a minimal length. For each query sequence a vector of features is obtained. The features consist in the number and average length of the relevant subsequences shared with each of the protein families. Classification is performed by combining these features in a Bayes Classifier. The combination of these characteristics results in a multi-class and multi-domain method that is exempt of data transformation and background knowledge. We illustrate the performance of our method using three collections of protein datasets. The performed tests showed that the method has an equivalent performance to state of the art methods in protein classification.

Data Mining

2005

P. Gabriel Ferreira, P. Azevedo, R.Alves, O.Belo, ***"A Hybrid Method to Discover Inter-Transactional Rules"***, in Proceedings of X Jornadas sobre Ingeniería del Software y Bases de Datos, JISBD'2005, (Granada, Spain, September 2005).

Classical association rules are by nature intra-transaction based, i.e., the associations occur among the items within the same transaction. Further, those transactions are usually associated with a dimensional context which is typically ignored. This prevents that patterns, occurring along this dimension, are discovered. We introduce a method for the extraction and analysis of inter-transactional patterns. These patterns allow discovering relationships that occur among transactions within a dimensional frame. The method consists in the combination of association and sequence mining. First, association mining is applied to transform a dimensional transactional database into a sequence database. The latter is then mined in order to obtain frequent sequence patterns. Sequence patterns are augmented with additional distance (among events) information. Finally, sequence rules fulfilling user's constraints (with relation to the measures of confidence, lift and others) are extracted. These rules have applications in a wide range of domains like stock market, weather databases or in retail market pricing strategies.

2004

R.Alves, O.Belo, F.Calvancati & P. Ferreira, ***"Clickstreams, the basis to establish user navigation patterns on websites"***, in Proceedings of 5th International Conference on Data Mining 2004, (Malaga, Spain, August 2004).

Collecting and mining clickstream data from e-commerce sites has become increasingly important for marketing, advertising, and traffic analysis activities. Organizations are promoting many initiatives concerning user's navigation pattern discovering, in order to implement better sites, more functional and close to customers' needs. Basically, the main idea is to provide more quality of attendance in their sites, and, consequently, get more profitability. However, clickstream processing is not a simple task. The sequences of clicks are very difficult to handle using conventional techniques, essentially, due to their diversity and nature. They include a lot of aspects that reveal the multidimensional perspective of web data. OLAP technology provides today means and techniques to represent, store and analyse such kind of multidimensional data. However, it does not offer discovery driven analysis to support traversal pattern identification processes on web sites. Mining traversal patterns techniques can be applied in conjunction with OLAP as an integrated alternative for understanding those particular sequences of clicks. In this paper we present an integrated OLAP and mining approach specially conceived for exploring user navigation patterns based on clickstreams. We also describe the multidimensional structure provided for modelling click sequences and the OLAP operations and mining techniques that can be pushed over data cubes to bring up

navigation patterns.

Secção D – Enquadramento

Enquadramento Científico

O objectivo deste plano de Dissertação de Doutoramento é o de contribuir para o melhoria e desenvolvimento do processo de automatização da descoberta de conhecimento em colecções de dados bioquímicos, em concreto na descoberta e extracção de padrões sequenciais. A metodologia a seguir para atingir este objectivo será a de estudo, selecção, aplicação e desenvolvimento de técnicas de computação apropriadas, nomeadamente de Data Mining e Machine Learning. Tendo sempre em conta por um lado, as características dos dados biológicos e por outro as particularidades de todo o processo de Descoberta de Conhecimento automático. Os problemas a tratar são computacionalmente pesados requerendo estruras de dados e algoritmos eficientes.

No final deste projecto espera-se contribuir para esta área de investigação com o aperfeiçoamento de metodologias já existentes e desenvolvimento de novas metodologias de forma a poder oferecer novas soluções para os desafios colocados pela progressiva transformação da biologia numa ciência de enormes quantidades de dados. Este trabalho poderá ter impacto em outras áreas onde haja necessidade de se usar pesquisa de padrões, com especial relevância para a biotecnologia, farmacologia e medicina.

Motivação

Actualmente a tarefa de extracção de padrões sequenciais é abordado sobre duas perspectivas distintas. Por um lado a comunidade de Data Mining, em que a investigação é realizada sobretudo por informáticos e cientistas da computação. Esta comunidade tem uma abordagem ao problema mais virada para a eficiência dos algoritmos e das estruturas de dados utilizadas. Por outro lado a comunidade de Bioinformática, cuja a investigação nesta área é feita essencialmente por bioquímicos que têm um problema biológico em concreto para resolver. Neste caso o problema está bem definido e o objectivo final é a sua resolução. A informática serve apenas como um meio. Ambas as abordagens têm as suas vantagens/desvantagens e raramente se intersectam nos seus pontos de vistas. Será assim muito interessante e relevante estudar em profundidades as duas perspectivas e tentar a sua combinação de forma a que possam resolver de forma mais eficiente o problema da extracção automática de padrões sequenciais em dados biológicos. Dada as diferentes características das sequências de ADN e de proteínas, decidiu-se focar o estudo no caso particular das proteínas.

De forma a demonstrar o interesse prático deste tipo de análise e nomeadamente da utilidade dos padrões obtidos, pretende-se igualmente estudar aplicações para esses mesmos padrões. Exemplos de aplicações são: classificação, clustering, anotação. Durante esta tese deu-se especial atenção à tarefa de classificação de sequências de proteínas.

Dada a natureza combinatória do processo de mineração, o número resultante de padrões é muitas das vezes elevado, o que torna a análise dos mesmos extremamente difícil. De forma a facilitar esta análise é necessário aplicar medidas que permitam pontuar os padrões de acordo com a relevância estatística e biológica dos mesmos. Verifica-se que diferentes autores propõe diferentes medidas e que actualmente não existe um estudo que permita verificar a exactidão de cada uma destas medidas na pontuação dos padrões. Neste sentido propõe-se nesta tese um estudo comparativo de diferentes medidas de pontuação de padrões e respectiva avaliação dessas medidas na descoberta de padrões biologicamente relevantes.

Os algoritmos estudados e desenvolvidos têm muitas das vezes aplicações que vão para além da Bioinformática, podendo ser aplicados directamente em diversos domínios. Desta forma, sempre que possível estuda-se a aplicação desses mesmos algoritmos, de forma isolada ou combinada com outros algoritmos, em diferentes áreas de aplicação. Isto irá permitir uma maior visibilidade dos algoritmos propostos bem como um maior reforço na demonstração da utilidade dos mesmos.

Objectivos Detalhados

- Estudo e Análise de algoritmos de mineração de padrões sequenciais;

Neste ponto pretende-se fazer um estudo do "estado-da-arte" nesta área de investigação. O estudo pressupõe o levantamento exaustivo de todos os algoritmos propostos na literatura, bem como as características, vantagens e desvantagens de cada um dos métodos/algoritmos. Igualmente pretende-se fazer uma análise dos casos de sucesso da aplicação dos algoritmos e das características dos problemas para os quais cada um dos algoritmos são mais adequados.

- Análise comparativa dos algoritmos provenientes da comunidade de Data Mining e da comunidade Bioinformática, tendo em conta a sua adequação aos dados biomoleculares (sequências de proteínas/dna) ;

Neste ponto da investigação, o objectivo é focar no estudo dos algoritmos para o caso particular da mineração de dados biológicos, mais em concreto para as bases de dados das proteínas. Efectuar um levantamento de quais os problemas propostos pela comunidade de Bioinformática que ainda se encontram em aberto, e que portanto podem ainda ser motivo de aperfeiçoamento. Estudar como os técnicas propostas pela comunidade de Data Mining podem ser úteis para resolver estes problemas.

- Desenvolvimento e optimização de um algoritmo de mineração de padrões sequenciais em base de dados de proteínas e que suporte as várias restrições e propriedades necessárias a este tipo de mineração;

Após efectuado o estudo realizado nos dois pontos anteriores, verificou-se a necessidade de um algoritmo que combinasse e integrasse os diferentes tipos de

restrições ao processo de mineração de padrões, bem como a introdução de outras características que são indispensáveis sob o ponto de vista da biologia para obter padrões efectivamente interessantes. Tal algoritmo surge como a combinação de novas técnicas e de técnicas anteriormente propostas em outros trabalhos e que demostraram serem de grande eficiência. Além disso o algoritmo é único, na medida que permite extrair dois tipos de padrões, *rígidos* e *flexíveis*. Estes padrões podem ser utilizados para diferentes tipos de análise dos dados. Para provar a eficiência do algoritmo é feita um estudo comparativo com dois dos algoritmos (um para cada tipo de padrão) mais avançados. Este estudo demonstra que o algoritmo proposto possui um eficiência superior a estes algoritmos em quase todas as situações.

- Estudo da relação entre a relevância estatística e biológica dos padrões sequenciais;

Neste ponto prentende-se estudar o impacto das diferentes medidas de relevância estatística/informação para a pontuação. Estas medidas tem um grande impacto na determinação de quais os padrões mais relevantes a analisar, visto que a quantidade de padrões resultantes do processo de mineração é sempre muito grande.

- Estudo de casos de aplicação dos padrões sequenciais, nomeadamente em tarefas de classificação de sequências;

Neste ponto estuda-se a aplicação directa dos padrões à tarefa de classificação de sequências de proteínas. A partir de uma base de dados de proteínas previamente classificados, a sequência a classficar é analisada em termos de padrões relevantes que possuem com cada uma destas famílias. Um padrão é relevante quando ocorre em um determinado número de outras sequências e quando o seu tamanho é maior ou igual a um valor definido pelo utilizador. Estes padrões são depois combinados num classificador Bayesiano que retorna um valor de similaridade da sequência em relação a cada família de proteínas.

- Aplicação, mediante uma eventual transformação dos dados, dos algoritmos de mineração a dados de séries temporais;

As séries temporais caracterizam-se por serem um conjunto ordenado de valores naturais. Neste ponto propomos aplicação do algoritmo de mineração a séries temporais. Para tal é efectuada uma redução de complexidade/dimensionalidade das séries temporais através de operações de *agregação* e *simbolização*. Posteriormente o algoritmo proposto no terceiro ponto, é adaptado de forma a permitir a extracção de padrões sequenciais em séries temporais.

- Estudo da aplicação e da combinação dos algoritmos de mineração de padrões sequenciais com outros algoritmos, nomeadamente de associação, para a extracção de padrões de natureza temporal;

Neste ponto desenvolveu-se um novo método que têm por base a combinação do algoritmo gIL com um algoritmo de associação. Esta combinação irá permitir

encontrar relações que são recorrentes ao longo da linha do tempo, uma vez que surgem de forma ordenada sob a forma de padrão sequencial. Estes algoritmos têm uma grande número de aplicações, sendo que as bases de dados dos valores de bolsa são normalmente um caso clássico de aplicação.

Trabalhos Alternativos e Bibliografia Principal

Classificação

No problema de classificação de sequências três tipos de métodos podem ser distuiguidos:

- *Classificadores Directos*, que exploram a natura sequencial dos dados comparando a similaridade entre sequências. Exemplo destes tipos de classificadores é o métod K-NN. Neste método a classe de cada uma das K sequências mais similares com a sequência a classificar são utilizadas para votar numa decisão. A similaridade entre sequências pode ser medida através de métodos como o FASTA [1] ou o BLAST [2].

- *Classificadores Baseados em Características*, que funcionam através da extracção e modelação de um conjunto de características das sequências e adaptando essas características com as técnicas tradicionais, como as árvores de decisão, baseadas em regras ou SVM [3,4,5,6,7]

- *Modelos Probabilísticos*, funcionam simulando a família de sequências em consideração e emitindo um valor de probabilidade que uma dada sequência a classificar irá ter de ser emitida pela família em questão. Estes classificadores são: cadeias de Markov de ordem K [8], Hidden Markon Models [9] e Probabilistic Sufix Trees [10].

Referências

[1] Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schae_er, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman. Gapped BLAST and PSIBLAST: a new generation of protein database search programs. Nucleic Acids Research, 25:3389{3402, 1997.

[2] R.W. Pearson and D.J. Lipman. Improved tools for biological sequence comparison. Proceedings Natl. Academy Sciences USA, 5:2444{2448, 1998.

[3] Daniel Kudenko and Haym Hirsh. Feature generation for sequence categorization.
In AAAI/IAAI, pages 733{738, 1998.

[4] Neal Lesh, Mohammed J. Zaki, and Mitsunori Ogihara. Mining features for sequence classi_cation. In Proceedings of the _fth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 342{346. ACM Press, 1999.

[5] A. Ben-Hur and D. Brutlag. Remote homology detection:a motif based approach. Bioinformatics, 19(1):26{33, 2003.

[6] A. Ben-Hur and D. Brutlag. Sequence motifs: highly predictive features of protein function. In Proceeding of Workshop on Feature Selection, NIPS – Neural Information Processing Systems, December 2003.

[7] N.M. Zaki, R.M. Ilias, and S. Derus. A comparative analysis of protein homology detection methods. Journal of Theoretics, 5.

[8] Durbin R. and Eddy S. R. Biological sequence analysis: Probabilistic models of proteins and nucleic acids. Cambridge University Press, 1998.

[9] Mian Sojlander Krogh, Brown and Haussler. Hidden markov models in computational biology: applications to protein modeling. Journal of Molecular Biology, (235):1501{1531, 1994.

[10] Bejerano G. and Yona G. Modeling protein families using probabilistic sufix trees.
In ACM press, editor, In the proceedings of RECOMB1999, pages 15{24, 1999.

Mineração de Padrões Sequenciais

Em relação à mineração de padrões sequenciais, tal como já foi referido anteriormente existem duas abordagens sobre este tema de investigação. No que diz respeito à comunidade de Data Mining e considerando o tipo de padrões extraídos, o gIL enquadra-se na categoria dos algoritmos: AprioriAll [1], Spade [2], PrefixSpan [3] e Spam [4], uma vez que todos estes algoritmos extraem padrões flexíveis de com "gaps" de tamanho arbitrário. No entanto é de salientar que o gIL trabalha sobre sequências lineares enquanto que os algoritmos mencionados trabalham sobre sequências transaccionais.

Do ponto de vista da Bioinformática, o algoritmo Teiresias [5] reporta padrões rígidos da classe de padrões fechados. Os padrões rígidos parecem indicar regiões das proteínas melhor conservados. Outro algoritmo de salientar é o ToMMS [6] que é um algoritmo que funciona sob uma abordagem "Top-Down" e reporta padrões máximos sem "gaps". Ambos os algoritmos funcionam com conjuntos de substituição, no entanto ToMMS não inclui restrições e no Teiresias apenas a restrição de "gap" máximo está contemplada. O algoritmo gIL também suporta conjuntos de substituição e suporta todas as restrições: "gap" máximo e mínimo, tamanho de janela, e eventos.

Referências

[1] Rakesh Agrawal and Ramakrishnan Srikant, Mining sequential patterns, Eleventh International Conference on Data Engineering (Taipei, Taiwan) (Philip S. Yu and Arbee S. P. Chen, eds.), IEEE Computer Society Press, 1995,
pp. 3-14.

[2] Mohammed J. Zaki, Spade: An efficient algorithm for mining frequent sequences, Machine Learning 42 (2001), no. 1-2, 31-60.

[3] J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M.-C. Hsu, PrefixSpan mining sequential patterns eficiently by prefix projected pattern growth, Proceedings Int. Conf. Data Engineering (ICDE'01), Heidelberg, Germany, April 2001, 2001, pp. 215-226.

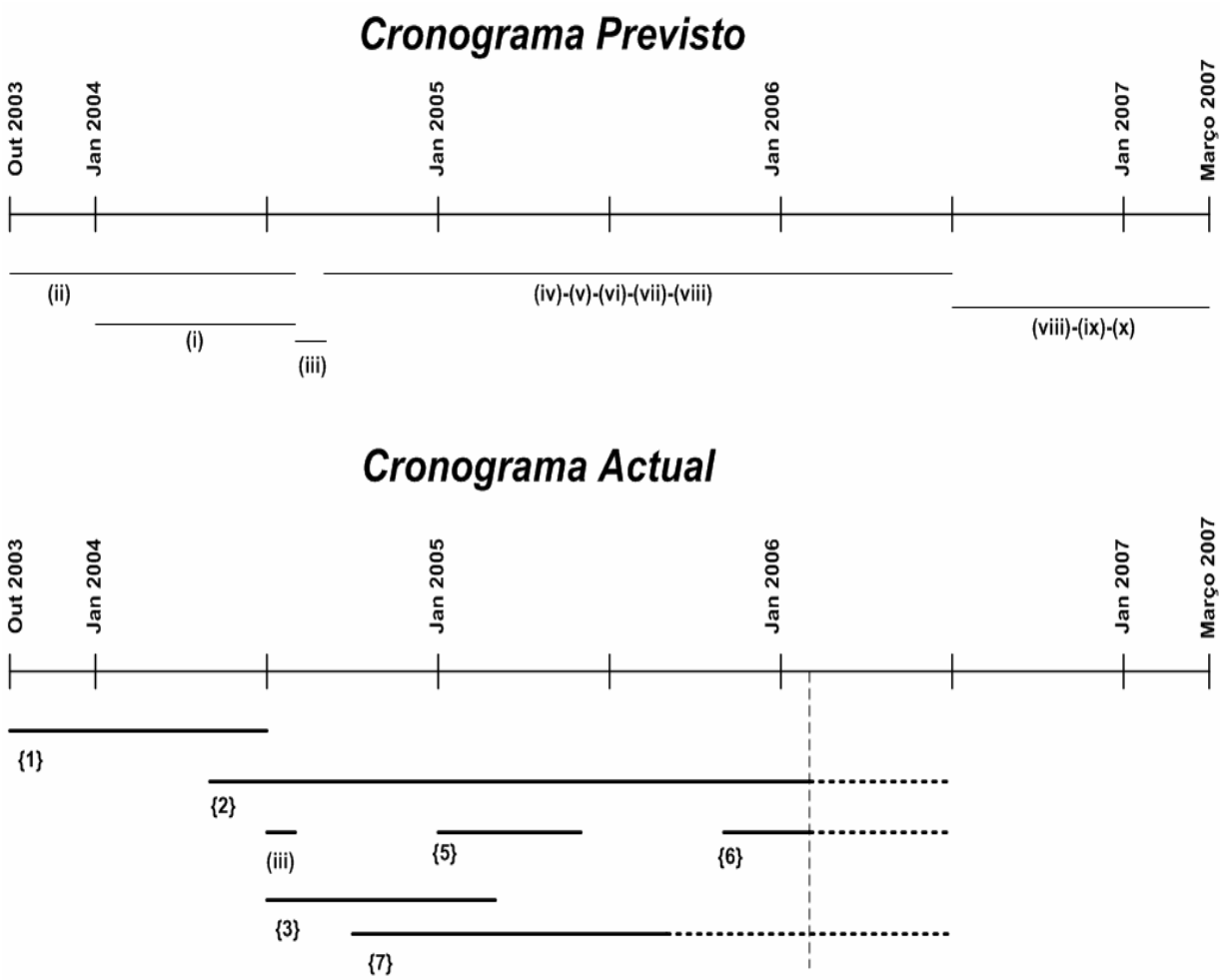
[4] J. Ayres, J. Flannick, J. Gehrke, and T. Yiu, Sequential pattern mining using a bitmap representation, Proceedings
of the 8th International Conference of Knowledge Discovery and Data Mining SIGKDD, S. Francisco, July 2002., 2002, pp. 429-435.

[5] - A.Floratos I. Rigoutsos, Combinatorial pattern discovery in biological sequences: the teiresias algorithm, Bioinformatics 1 (1998), no. 14.

[6] Martin Ester and Xiang Zhang, A top-down method for mining most speci_c frequent patterns in biological sequence data, Proceedings of the SIAM International Conference on Data Mining (SDM 2004), Lake Buena Vista, FL, USA, 2004, 2004.

Secção E – Desenvolvimento

Macro-planeamento das Actividades



Cronograma das etapas realizadas até ao momento. Os números que descrevem cada etapa são identificadores dos objectivos propostos na secção B.3 deste relatório.

Recursos Necessários

Para a realização das tarefas propostas é necessário uma máquina com uma capacidade de processamento actual e algum software específico para desenvolvimento do software protótipo. Tendo em conta que os algoritmos propostos devem correr em desktops, a máquina para desenvolvimento do algoritmo deve ter uma capacidade de processamento idêntica. Isto irá dar valores exactos em termos de eficiência do algoritmo em máquinas desktop. O protótipo desenvolvido irá ser disponibilizado para acesso livre, desta forma as ferramentas a utilizar no seu desenvolvimento também irão ser de acesso livre GNU/GPL, nomeadamente um compilador de código C/C++ (cygwin) e um editor de texto. As bases de dados de proteínas necessárias para testar o algoritmo estão disponíveis gratuitamente e de acesso livre na Internet (ex: Pfam, Prosite, etc.). De uma maneira geral não se prevê dificuldades em conseguir reunir os meios para a realização das tarefas propostas.

Recursos Disponibilizados

Os recursos disponibilizados acabaram por ir de encontro as necessidades acima indicadas. No entanto foi inicialmente indicado ao aluno que não haveria grande disponibilidade de uma máquina actual. Pelo que no período inicial esta falha teve que ser colmatada com a utilização do computador portátil do aluno. Mais tarde (cerca de ano e meio após o início do doutoramento) o aluno passou a ter acesso a uma máquina actualizada, mas que foi adquirida no âmbito de um projecto de investigação em que o aluno estava envolvido. Em termos de software não houve qualquer tipo de problemas .

Secção F – Avaliação

Análise Comparativa

Em termos gerais pode dizer-se que o trabalho desenvolvido está ao nível dos trabalhos mais recentes realizados nas áreas de investigação a que dizem respeito.

Em relação ao trabalho desenvolvido no âmbito do objectivo 3, verificou-se através de uma avaliação experimental que o algoritmo possui uma eficiência igual ou superior aos algoritmos considerados até à data como mais eficientes nas respectivas tarefas de mineração.

No algoritmo proposto para classificação de sequências, objectivo 5, a precisão do algoritmo é comparável e em alguns casos superior à dos algoritmos mais recentes de classificação de proteínas. Além disso o algoritmo proposto revela-se muito mais simples de concretizar.

Auto-Avaliação do Trabalho Realizado

Fazendo uma análise global do trabalho realizado até à data o aluno considera que à data de escrita deste relatório se encontram realizados de 60 a 70% dos objectivos estratégicos propostos. Actualmente a investigação tem-se centrado no estudo da aplicação dos algoritmos de extracção de padrões sequenciais a séries temporais, com particular ênfase na mineração de dados resultantes da simulação do processo de folding/unfolding de proteínas. Outros dois pontos que faltam ainda abordar com mais profundidade neste doutoramento são:

Aplicação do método que combina os algoritmos de mineração de associação/sequências, a um caso particular de dados biológicos. Até o momento este método foi apenas aplicado a dados gerados sinteticamente. Supõe-se que este método possa vir a ser de grande utilidade na análise de dados biológicos.

O segundo ponto a abordar é o estudo de medidas de relevância estatística e informativa dos padrões sequenciais e a sua relação com o significado biológico dos dados. Este estudo pode ter bastante impacto nesta área de investigação uma vez que não existe actualmente na literatura um outro estudo semelhante.

Auto-Avaliação da Documentação Produzida

Dos trabalhos já publicados e acima enumerados, o primeiro descreve as actividades de investigação realizadas no âmbito do tópico 3 dos objectivos propostos. O segundo artigo publicado descreve o tópico 5. Estes dois tópicos de investigação encontram-se concluídos uma vez que se julga terem sido atingidos os objectivos propostos. A terceira e quarta publicação da lista referem-se ao objectivo 7. Adicionalmente, existem 2 artigos escritos sobre a forma de tutoriais. Estes artigos têm sido objecto de constantes alterações/incrementos e ainda não foram publicados. Os artigos incidem sob o tópico 1 e 2 dos objectivos propostos. Actualmente encontra-se em fase de escrita uma artigo que resume as actividades de investigação realizadas no âmbito do objectivo 6. Pode-se admitir que estas quatro publicações reflectem cerca de 50% do trabalho desenvolvido pelo aluno.

Secção G Dificuldades

Dificuldades Técnico-Científicas

Estando a investigação a ser feita numa área interface a várias disciplinas, existiu sempre um dificuldade inerente à falta da formação a uma das áreas. Neste tema de investigação existe uma intersecção da disciplina da computação, com a biologia. Sendo a formação do aluno em informática, alguns assuntos directamente relacionados com a biologia surgiram como dificuldade. Após uma fase inicial em que o aluno se dedicou ao estudo dos princípios básicos da biologia molecular estes conceitos passaram a ser analisados com mais algum "à vontade". As maiores dificuldades surgiram aquando do momento de encontrar problemas biológicos e casos de teste muito específicos que demonstrassem à comunidade da biologia a aplicabilidade e a utilidade dos algoritmos desenvolvidos. Houve sempre que possível a interacção com alguns biólogos, mas devido ao facto de esta área não ser totalmente uma especialidade das pessoas em questão, apenas contribuições limitadas foram sendo dadas ao trabalho. O aluno utilizou sempre como referência para a sua investigação outros trabalhos já publicados,

recriando as experiências neles indicadas para a avaliação experimental dos algoritmos propostos. Em geral pode dizer-se que esta limitação de apoio de especialistas na área bioquímica foi em parte ultrapassada e que não levou ao desvio dos objectivos inicialmente propostos. No entanto é opinião do aluno de que um maior apoio neste sentido, teria levado à produção de um trabalho ainda mais bem fundamentado sobre o ponto de vista aplicacional.

Outras Dificuldades

Neste aspecto o aluno não têm nada a apontar, salientando ainda o bom ambiente e o espírito de entre-ajuda que encontrou e têm encontrado neste departamento com colegas de doutoramento, orientador, professores e funcionários.