PERL script para busca de erros e não conformidades em um banco de dados biológico público

Flávia G. Silva[1], Kátia P. Lopes[2], Sandro R. Dias[1]

[1]Departamento de Sistemas de Informação – Faculdade Anhanguera de Belo Horizonte
Av. dos Andradas, 436, Centro, Belo Horizonte - MG, Brasil

[2]Programa de Pós-Graduação em Bioinformática
Universidade Federal do Paraná (UFPR) – Curitiba-PR, Brasil

{flagomesbh,katiaplopes}@gmail.com, sandro.dias@aedu.com

Especificamente para este trabalho, foi utilizado o Banco de dados de proteínas Protein Data Bank (PDB), disponível em: <www.rcsb.org>. Esta base de dados é amplamente utilizada por pesquisadores da área de Bioinformática e armazena informações tridimensionais (3D) de estruturas de proteínas, ácidos nucleicos e conjuntos complexos, estrutura preliminar e secundária das proteínas, bem como ângulos e distâncias entre os átomos. Esses dados são armazenados em um conjunto de arquivos contendo uma nomenclatura padronizada, chamados de *flat files*. Porém, nem todos os arquivos seguem o padrão proposto ou erros passam despercebidos pelo pesquisador, fazendo com que análises futuras tenham que passar por um processo manual de curadoria. Primeiramente, a partir do trabalho de Dias & Nagem (2009) foram executados scripts nas linguagens php e shell no banco PDB. Durante o processo de recolher informações referentes às identificações dos métodos utilizados para gerar os arquivos, informações das identificações dos arquivos, e dados referentes ao enxofre Gama, foi possível verificar que alguns arquivos do PDB apresentam os seguintes problemas: 1) Os átomos não estão com numeração consecutiva indicando que nem todos os átomos foram resolvidos, ou seja, nem todos os átomos que compõem a proteína estão presentes no arquivo. 2) O número de resíduo também não é consecutivo, assim, o pesquisador não tem como identificar se foi um erro de numeração apenas, ou se nem todos os átomos foram resolvidos para determinada proteína. 3) O número de resíduo contém, além do número uma letra, ferramentas como CHASA, PyRosetta, dentre outras que trabalham com modelagem de proteínas, não conseguem manipular/visualizar/pesquisar arquivos com esta falha. 4) Muitos resíduos com o mesmo "identificador único", o que dificulta o processo de manipulação do arquivo, principalmente para tratamento dos dados para inserção em um banco de dados relacional. Portanto, foi desenvolvido um algoritmo em PERL script que percorre um diretório com arquivos .pdb à procura dos problemas citados, baseando-se na documentação do próprio PDB disponível em sua página na internet. Como exemplo, o arquivo 3GBN.pdb referente a uma proteína do vírus H1N1 demonstrou alguns destes problemas: Na linha 6367 a numeração dos resíduos é alterada de 124 para 141, onde deveria ser consecutivo. O intervalo de 4284 a 4291 (linha 5751) apresenta uma não-conformidade: a numeração dos resíduos PRO (prolina) é a mesma da ILE (isoleucina), incluindo apenas a letra A além do número 52. Em um teste preliminar, portanto, o algoritmo proposto atendeu de forma eficiente a demanda da curadoria, tendo gasto um tempo de 2min14s para 1023 arquivos pdb e espera-se dessa forma, contribuir positivamente com as pesquisas realizadas na base de dados do PDB.