BIOSOM: Metodologia para identificação de sinônimos de genes utilizando *Self-Organizing Maps*.

Kelly Rafaela Otemaier[1], Maria B. R. Steffens[1, 2], Roberto T. Raittz[1], Jeroniza N. Marchaukoski[1]

[1] Programa de Pós Graduação em Bioinformática - Universidade Federal do Paraná(UFPR)
[2] Departamento de Bioquímica e Biologia Molecular - Universidade Federal do Paraná (UFPR)
rafaela.otemaier@gmail.com,{steffens, raittz, jeroniza}@ufpr.br,

A preocupação com a nomenclatura de genes existe desde que se iniciaram as anotações gênicas e o problema ainda está longe de ser solucionado. Existem diversas diretrizes para nomenclatura de genes, mas elas não são rigorosamente aplicadas à atribuição de nomes aos genes recém-identificados, gerando assim, inúmeras maneiras de nomear um mesmo gene. Nomenclaturas uniformizadas melhoram a documentação das sequências gênicas depositadas nos bancos de dados públicos e facilitam os processos de análise de dados e anotações de novas sequências. Atualmente é comum encontrar genes com a mesma função e nomes diferentes ou variações de um nome para o mesmo gene. Buscando a padronização da nomenclatura gênica, foi utilizada a técnica de *Self-Organizing Maps* (SOM) em conjunto com a de *Matrix-U* para identificar sinônimos de genes através de agrupamentos. A rede neural artificial SOM é estruturada a partir do aprendizado competitivo para identificar as semelhanças das características dos nomes de genes e a relação de vizinhança entre estes nomes. A técnica de *Matrix-U* é aplicada para a identificação dos agrupamentos gerados pela rede neural SOM. Neste trabalho, foram selecionados do banco de dados de genomas completos do Genbank/NCBI dez genes distintos. A seguir estes genes foram submetidos ao Blast, realizando alinhamentos e gerando, cada gene, um conjunto composto por outros 100 genes. Os resultados obtidos dos alinhamentos foram normalizados, passados por uma linha de corte de *E-value* < 1e-10 e finalmente inseridos na rede neural SOM através do software Matlab. A técnica de *Matrix-U* foi aplicada aos resultados da rede neural SOM para detectar os agrupamentos de dados baseados nas características sintáticas dos nomes e nos valores de referência do Blast. Como resultado o erro topográfico médio foi de 0,24 e o erro médio de quantização foi de 0,16, o que indica que os mapas gerados estão bem ajustados aos vetores de entrada. Os resultados dos agrupamentos foram avaliados por especialistas na área biológica, que validaram os grupos gerados. Nos experimentos realizados, foi possível identificar em um dos grupos, um gene que não pertencia à mesma família dos demais, o que pode indicar um provável erro de anotação. Com essa indicação, poderiam ser realizados procedimentos laboratoriais que confirmassem a hipótese levantada. A metodologia desenvolvida foi é aplicável para descrever genes hipotéticos, *putative* e outros sem uma função descrita ou conhecida, podendo indicar uma possível função a estes genes após o agrupamento. Esta metodologia pode ser utilizada com qualquer sequência de aminoácido que gere um grupo de dados através de alinhamentos, onde possa ser submetidos à rede SOM.