GMR Genetics and Molecular Research
Online Journal
ISSN:1676-5680

Biosom: gene synonym analysis by self-organizing map

K.R. Otemaier[1], M.B.R. Steffens[1,2], R.T. Raittz[1], A. Brawerman[1] and J.N. Marchaukoski[1]

[1]Programa de Pós-Graduação em Bioinformática, Universidade Federal do Paraná, Curitiba, PR, Brasil
[2]Núcleo de Fixação Biológica de Nitrogênio, Departamento de Bioquímica e Biologia Molecular, Universidade Federal do Paraná, Curitiba, PR, Brasil

Corresponding author: R.T. Raittz
E-mail: raittz@gmail.com

Genet. Mol. Res. 14 (1): 1461-1468 (2015)
Received January 9, 2014
Accepted October 29, 2014
Published February 20, 2015
DOI http://dx.doi.org/10.4238/2015.February.20.1

ABSTRACT. There are several guidelines for gene nomenclature, but they are not always applied to the names of newly identified genes. The lack of standardization in naming genes generates inconsistent databases with errors such as genes with the same function and different names, genes with different functions and the same name, and use of an abbreviated name. This paper presents a methodology for predicting synonyms in a given gene nomenclature, thereby detecting and minimizing naming redundancy and inconsistency and facilitating the annotation of new genes and data mining in public databases. To identify gene synonyms, i.e., gene ambiguity, the methodology proposed begins by grouping genes according to their names using a Kohonen self-organizing map artificial neural network. Afterwards, it identifies the groups generated employing the Matrix-U technique. The employment of such techniques allows one to infer the synonyms of genes, to predict probable hypothetical gene names and to point out possible errors in

a database record. Many mistakes related to gene nomenclature were detected in this research, demonstrating the importance of predicting synonyms. The methodology developed is applicable for describing hypothetical, putative and other types of genes without a known function. Moreover, it can also indicate a possible function for genes after grouping them.

Key words: Gene nomenclature; Gene ambiguity; Kohonen; Gene synonym prediction; Self-organizing map; Matrix-U

INTRODUCTION

The scientific community is currently witnessing an exponential increase in new organisms. Moreover, this rate is tending to grow even further due to the price reduction of automated DNA sequencing and the importance of these new discoveries. Therefore, the use of a standard to choose and represent genes is of fundamental importance to store and recover this great amount of information (Lopes and Cruz, 2011).

There has been concern over the nomenclature of genes and proteins ever since gene annotation began at the time of Gregor Mendel in 1860. A century later, in the 1960s, the study and discussions of the first problems related to gene nomenclature began (Demerec et al., 1966).

Various efforts were made towards an agreement on gene and protein nomenclature, for instance, the Demerec system (Demerec et al., 1966) concerning nomenclature for bacterial genes, HUGO Gene Nomenclature Committee (HGNC) regarding human gene nomenclature (Eyre et al., 2006) and UniProt Consortium towards protein nomenclature (Consortium, 2009).

There are many guidelines for gene nomenclature; however, they are not rigorously applied to newly identified genes. Normally, researchers are free to define and assign the name they feel best suited to their discoveries. As a result, there may be innumerous ways of naming the same gene, such as full name, symbol, and synonym (Tsuruoka et al., 2007; Huang et al., 2010).

It is common to find in public databases and in the literature genes with the same function but different names or name variations of the same gene, thus creating doubts and confusion in adopting a name for a new gene.

One way of reducing gene nomenclature ambiguity may be to identify synonyms by grouping the different ways in which a gene is annotated (Tsuruoka et al., 2007; Huang et al., 2010). The identification of synonyms provides an improvement in the documentation of gene sequences stored in public databases, which may ease the processes of database analysis and annotations of new gene sequences, especially the automated ones (Liu et al., 2006).

This study presents a new methodology based on the self-organizing map (SOM) artificial neural network (Kohonen, 1990), called BIOSOM. The goal was to form gene groups using the SOM network and identify gene nomenclature ambiguities employing the Matrix-U technique, thus reducing redundancy and inconsistency in gene nomenclature.

To validate the methodology proposed, experiments were performed to evaluate the efficiency of the SOM network in identifying gene ambiguity. This technique of visualization is rather interesting, since it does not require the number of groups being formed to be

known. Therefore, the groups are formed through characteristics indicated by the neural network.

MATERIAL AND METHODS

The set of full bacterial genome genes employed for validation of the methodology proposed was obtained from the National Center for Biotechnology Information (NCBI) on June 3, 2011. Table 1 lists the genes arbitrarily selected to be used in ten experiments, each containing a group of data composed of 100 genes. Figure 1 shows the methodology workflow developed.

-Step A (Figure 1A): An analyzer software for NR (non-redundant - NCBI) files was developed in Java language. The analyzer is responsible for reading the NR file, interpreting the patterns contained in it and later storing them in a local database.

-Step B (Figure 1B): A local database was created using the PostgreSql Database Management System, version 8. This database stores data extracted by the NR analyzer software.

-Step C (Figure 1C): The amino acid sequences extracted for this study were subjected to the software BlastP (Huang et al., 2010), in which the sequences were aligned. Each of the ten sequences previously selected served as an entry (query) to the BlastP, thus generating ten distinct data sets containing 100 genes each. Figure 2 depicts this entire step.

-Step D (Figure 1D): Each data set generated went through a selection of characteristics, which extracted the high-scoring segment pair (Hsp) values – Hsp E-value, Hsp qseq (query sequence), Hsp hseq (subject sequence), Hsp hit-to (size sequence), Hsp positive (similarity alignment) and Hsp identity (identity alignment).

-Step E (Figure 1E): The cutoff E-value filter was applied (Kohonen, 1990; Pavy et al., 2005; Frech and Chen, 2010; Belda-ferre et al., 2011; Yi and Jung, 2011). This filter was used so that low value alignments would not interfere in the final grouping.

-Step F (Figure 1F): Normalization of gene naming. The ambiguity and small variations of names were both minimized. The techniques employed were the following:

1) Converting uppercase into lowercase (Cohen et al., 2002);

2) Removing hyphen (Bruijn and Martin, 2003; Fang et al., 2006);

3) Removing extra blank spaces (beginning, middle, end) (Fang et al., 2006);

4) Removing parentheses (Fang et al., 2006);

5) Converting Roman numerals to Arabic numbers (Bruijn and Martin, 2003).

-Step G (Figure 1G): The INREC technique was used to perform a dimensionality reduction (Souza, 1999).

-Step H (Figure 1H): Neural network training using SOM Toolbox (Mathworks, 2008).

-Step I (Figure 1I): Finally, after carrying out step H, the groups were identified using the Matrix-U technique (Ultsch, 1993).

RESULTS

Ten experiments were conducted using the methodology presented in Figures 1 and 2, each containing a group of data composed of 100 genes. The genes used in the validation were arbitrarily obtained using the BlastP software and are given in Table 1. The data presented in Table 1 were used to analyze and validate the groups of genes found.

Table 1. Amino acid sequence submitted to the BLASTP software.

Gene name	Gi identificator
Argininosuccinate lyase	23335287
Abc transporter atp-binding protein	15802782
Resolvase	9507569
Dna-binding response regulator phob s4	15640738
Ribonucleotide-diphosphate reductase subunit beta	16804410
Fructose-specific phosphotransferase system protein frvx	16131738
Pyridoxamine-phosphate oxidase protein	333905884
2-component transcriptional regulator	15803079
Ferritin-like protein	15802782
Response regulator receiver modulated metal dependent phosphohydrolase	15599975

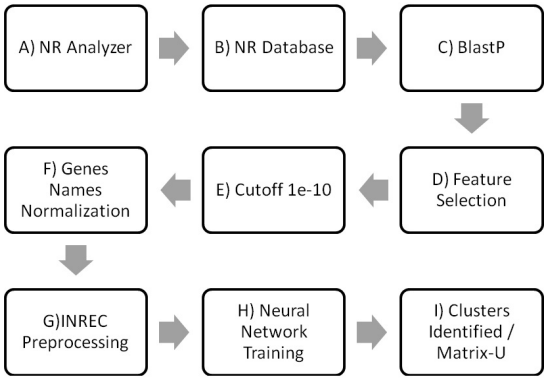

Figure 1. Methodology workflow. Nine steps to completely execute the proposed workflow and eliminate gene nomenclature ambiguity.

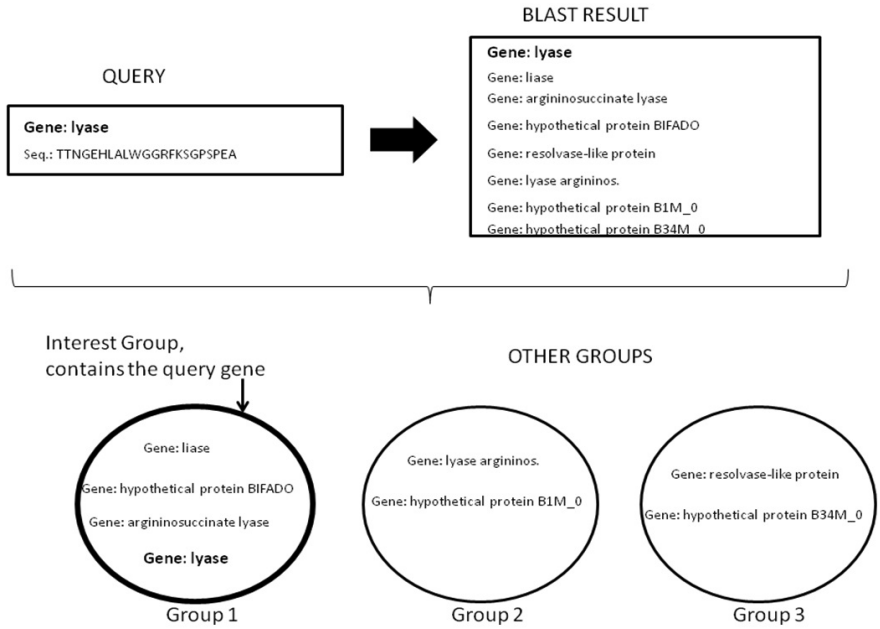

Figure 2. Blast results show ambiguity. Ambiguities in Blast results generating different options.

Figure 3 briefly elaborates the results obtained in the ten experiments conducted, where the x-axis represents the number of genes in each cluster, and the y-axis identifies the experiments. The colors are used to represent the cluster obtained in each experiment.

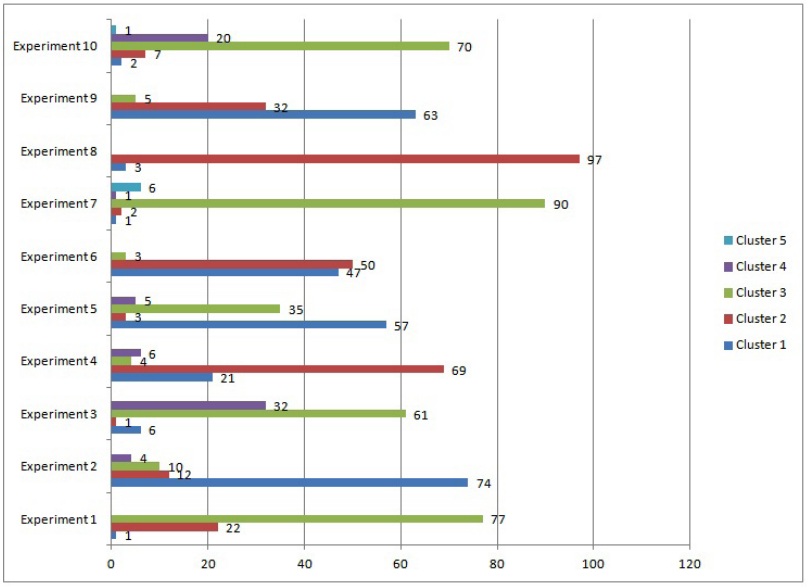

Figure 3. Genes clusters identified in each experiment.

Table 2 shows experiment 2 in more details. The results obtained in this experiment showed there could be several ambiguity occurrences to eliminate, namely names of genes belonging to the same family, names of genes belonging to different families, hypothetical gene names, generic gene names (e.g., only the family name, the domain name or even no names at all).

Experiment 2 (Table 2) indicated a probable annotation error. Cluster 4 included the pyridoxamine 5'-phosphate oxidase gene, where genes belonging to the ABC_tran (PF00005) family prevailed in this cluster. We conducted then a search in the PFAM (Protein FAMilies database) using the name of the pyridoxamine 5'-phosphate oxidase protein gene and found that this gene belongs to the Pyridox_oxidase (PF01243) family. Finally, we performed the alignment between the sequences of the ABC transporter ATP-binding protein and pyridoxamine 5'-phosphate oxidase protein genes, resulting in a percentage of identity of 85%. This also indicated that the gene in question should be annotated as an ABC transporter and not as a pyridoxamine 5'-phosphate oxidase protein gene, which discarded any possibility of the pyridoxamine 5'-phosphate oxidase protein gene having a connection with the ABC_tran (PF00005) family.

Figure 4 shows the quantization and topographical errors obtained in the experiments. When the quantization errors diminish, there is an increase of the topographical error. According to Kohonen (1990) and Yi and Jung (2011), the best map is the one that has the smallest quantization errors, since it would be more easily adjusted to the entry vectors. In this study, the standard parameters group was placed mostly in cases that showed quantization errors smaller than 0.2 (this value was inferred by the simple average of the quantization error variation obtained in the experiments).

Table 2. Four clusters found in one experiment.

Cluster	Gene name	Gene value	Family
1	Hypothetical protein lmo2580	01	-
	Hypothetical protein lin2725	01	-
	Hypothetical protein lin2471	01	-
	Hypothetical protein lmo2372	01	-
2	ABC transporter, ATP-binding protein	04	ABC_tran (PF00005)
	Putative hemin import ATP-binding protein hrta	02	ABC_tran (PF00005)
	Lipoprotein-releasing system ATP-binding protein lold	02	ABC_tran (PF00005)
	Hypothetical protein monocytfsl_07750	01	-
	ABC superfamily ATP binding cassette transporter, ABC protein	01	ABC_tran (PF00005)
3	ABC transporter ATP-binding protein	02	ABC_tran (PF00005)
	Lipoprotein-releasing system ATP-binding protein lold	01	ABC_tran (PF00005)
	ABC superfamily ATP binding cassette transporter, ABC protein	04	ABC_tran (PF00005)
	ABC transporter protein	03	ABC_tran (PF00005)
	ABC transporter related protein	03	ABC_tran (PF00005)
4	ABC-transporter ATP binding protein	49	ABC_tran (PF00005)
	Pyridoxamine 5'-phosphate oxidase protein	01	?
	Lipoprotein-releasing system ATP-binding protein lold	02	ABC_tran (PF00005)
	Hypothetical protein lmonf_01221	01	-
	ABC transporter family protein	01	ABC_tran (PF00005)
	Macrolide export ATP-binding/permease protein macb	02	ABC_tran (PF00005)
	Macrolide export ATP-binding/permease protein macb	01	ABC_tran (PF00005)
	Hypothetical protein lfark3_01742	01	-
	Hypothetical protein HMPREF0428_00848	01	-
	Lipoprotein releasing system, ATP-binding protein	01	ABC_tran (PF00005)
	Hypothetical protein HMPREF0433_01076	01	-
	Hypothetical protein LMRG_02687	01	ABC_tran (PF00005)
	Macrolide ABC transporter ATP-binding protein/permease	01	ABC_tran (PF00005)
	ABC-type antimicrobial peptide transport system, ATPase component	02	ABC_tran (PF00005)
	ABC transporter-like ATP-binding protein	01	ABC_tran (PF00005)
	Putative ABC transporter ATP-binding protein	02	ABC_tran (PF00005)
	ABC transporter	03	ABC_tran (PF00005)
	ABC transporter related protein	01	ABC_tran (PF00005)
	Hypothetica l protein CAT7_09005	01	-
	Phosphonate-transporting ATPase	01	ABC_tran (PF00005)
Sum		100	

Four distinct clusters in an experiment. In each group obtained it is possible observe the variation of gene names.

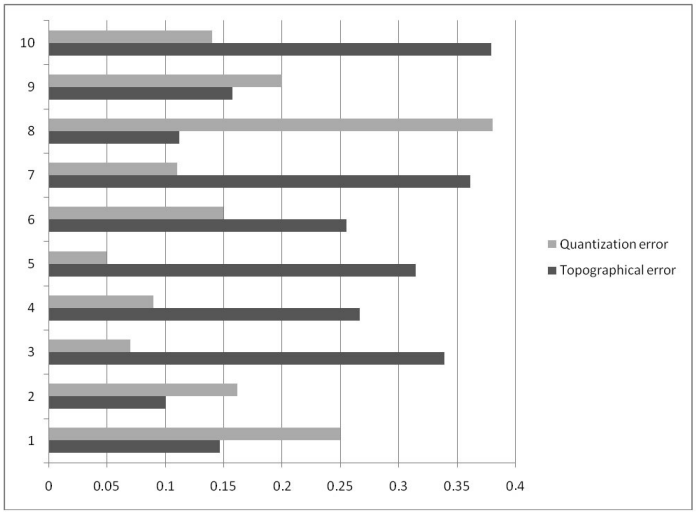

Figure 4. Quantization and topographic errors. Quantization and topographic error rate for each experiment performed.

To conclude this section, a comparison study between the main projects on gene nomenclature synonyms, namely Biothesaurus (Liu et al., 2006), GPSDB (Pillet et al., 2005), CD-HIT (Huang et al., 2010) and BIOSOM, is presented in Table 3.

Table 3. Comparing gene nomenclature synonyms projects.

Features	BioSom	Biothesaurus	GPSDB (Tsuruoka et al., 2007)	CD-HIT version 4.5.4 (Liu et al., 2006)
Sequence alignment	Yes	No	No	Yes
Syntactic form name gene	Yes	Yes	Yes	No
Database conection	Yes	Yes	Yes	No
Used data bases	A	B	C	D
Hypothetical genes include	Yes	No	No	-
Format file cluster	Text File (.txt)	Grouping preprocessed by gene name	Grouping preprocessed by gene name	Fasta file
Internet search	No	Yes	Yes	Yes
Error annotation identify	Yes	No	No	No

Comparison study among the main projects that propose somehow to reduce gene nomenclature ambiguity. A - GenBank, EMBL Data Library, DDBJ, NBRF PIR, Protein Research Foundation, SWISS-PROT, Brookhaven Protein Data Bank, Patents, NCBI Reference Sequence; B - UniProt, Swiss-Prot, TrEMBL, PIR-PSD, Entrez Gene, RefSeq and GenPept, MGD, SGD, RGD, FlyBase e WormBase, HUGO, EC enzyme nomenclature and OMIM; C - LocusLink, Swiss-Prot, GDB, HUGO, OMIM, MGD, RGD, Ratmap Flybase SGD, TAIR, WormBase, SubtiList and EcoGene; D - NCBI NR, Swissprot and PDBO. User may provide specific data base.

DISCUSSION

The lack of standardization in naming genes generates inconsistent databases with errors such as genes with the same function and different names, genes with different functions and the same name, and use of an abbreviated name (Tsuruoka et al., 2007; Huang et al., 2010).

The identification of synonyms provides an improvement in the documentation of gene sequences stored in public databases, which may ease the processes of database analysis and annotations of new gene sequences, especially the automated ones (Liu et al., 2006).

We presented a methodology based on the Kohonen self-organizing maps (SOM) artificial neural network (Kohonen, 1990) for predicting ambiguity in a given gene nomenclature, thereby detecting and minimizing naming redundancy and inconsistency.

We conducted ten experiments. Experiment 2 showed that there could be several ambiguity occurrences to eliminate, including names of genes belonging to the same family, names of genes belonging to different families, hypothetical gene names and generic gene names. In this experiment, cluster 4 showed the pyridoxamine 5'-phosphate oxidase protein gene in the family to be wrong, indicating a probable annotation error. With this indication, laboratory procedures can be carried out to confirm the hypothesis raised.

The methodology developed is applicable for describing hypothetic, putative and other genes without the need of a described or known function, and besides, it may actually indicate a possible function for these genes after grouping them.

The SOM artificial neural network employed presents the advantage of generating maps quickly, without the need of creating a great amount of iterations to obtain a good result.

Finally, viewing the maps through a Matrix-U allowed us to identify a quantity of clusters formed by means of a color matrix. The methodology described may also be used with any amino acid sequence that generates a group of data through alignments.

ACKNOWLEDGMENTS

Research supported by INCT de Fixação Biológica de Nitrogênio-Institutos Nacionais de Ciência e Tecnologia; Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). Ministério da Ciência e Tecnologia (MCT), and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

REFERENCES

Belda-Ferre P, Cabrera-Rubio R, Moya A and Mira A (2011). Mining virulence genes using metagenomics. *PLoS One* 6: e24975.

Bruijn BD and Martin J (2003). Finding Gene Function Using LitMiner. The Twelfth Text Retrieval Conference (TREC 2003), Gaithersburg, 486-494.

Cohen KB, Dolbey E, Acquaah-Mensah GK and Hunter L (2002). Contrast and Variability in Gene Names. Proceedings of the Workshop on Natural Language Processing in the Biomedical Domain, Philadelphia, 14-20.

Consortium (2009). The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res.* 37: D169-D174.

Demerec M, Adelberg EA, Clark AJ and Hartman PE (1966). A proposal for a uniform nomenclature in bacterial genetics. *Genetics* 54: 61-76.

Eyre TA, Ducluzeau F, Sneddon TP, Povey S, et al. (2006). The HUGO Gene Nomenclature Database, 2006 updates. *Nucleic Acids Res.* 34: D319-D321.

Fang HR, Jin Y, Kim JS and White PS (2006). Human Gene Name Normalization using Text Matching with Automatically Extracted Synonym Dictionaries. Proceedings of the BioNLP Workshop on Linking Natural Language Processing and Biology at HLT-NAACL 06, New York, 41-48.

Frech C and Chen N (2010). Genome-wide comparative gene family classification. *PLoS One* 5: e13409.

Huang Y, Niu BF, Gao Y, Fu LM, et al. (2010). CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* 26: 680-682.

Kohonen T (1990). The Self-Organizing Map. *Proc. IEEE* 78: 1464-1480.

Liu H, Hu ZZ, Zhang J and Wu C (2006). BioThesaurus: a web-based thesaurus of protein and gene names. *Bioinformatics* 22: 103-105.

Lopes HS and Cruz LM (2011). Computational Biology and Applied Bioinformatics. InTech Press, Rijeka.

Mathworks (2008). The Language Of Technical Computing. MATLAB. Available at [http://www.mathworks.com/]. Accessed January 18, 2012.

Pavy N, Paule C, Parsons L, Crow JA, et al. (2005). Generation, annotation, analysis and database integration of 16,500 white spruce EST clusters. *BMC Genomics* 6: 144.

Pillet V, Zehnder M, Seewald AK, Veuthey AL, et al. (2005). GPSDB: a new database for synonyms expansion of gene and protein names. *Bioinformatics* 21: 1743-1744.

Souza JA (1999). Recognizing Patterns Using Recursive Indexing. PhD thesis, Federal University of Santa Catarina, Santa Catarina.

Tsuruoka Y, McNaught J, Tsujii J and Ananiadou S (2007). Learning string similarity measures for gene/protein name dictionary look-up using logistic regression. *Bioinformatics* 23: 2768-2774.

Ultsch A (1993). Self-Organizing Neural Networks for Visualization and Classification. Information and Classification. Springer-Verlag Press, Heidelberg, 307-313.

Yi G and Jung J (2011). Algorithm for large-scale clustering across multiple genomes. *Bioinformation* 7: 251-256.