Bioinformatics, Interaction Network Analysis, and Neural Networks to Characterize Gene Expression of Radicular Cyst and Periapical Granuloma

Fabiano de Oliveira Poswar, MD, MS, Lucyana Conceição Farias, DDS, PhD,† Carlos Alberto de Carvalho Fraga, DDS, PhD,* Wilson Bambirra, Jr, DDS, MD,‡ Manoel Brito-Júnior, DDS, PhD,* Manoel Damião Sousa-Neto, DDS, MD, PhD,§ Sérgio Henrique Souza Santos, DDS, PhD,†‖ Alfredo Maurício Batista de Paula, DDS, PhD,* Marcos Flávio Silveira Vasconcelos D'Angelo, PhD,‖ and André Luiz Sena Guimarães, DDS, PhD**

Abstract

Introduction: Bioinformatics has emerged as an important tool to analyze the large amount of data generated by research in different diseases. In this study, gene expression for radicular cysts (RCs) and periapical granulomas (PGs) was characterized based on a leader gene approach. **Methods:** A validated bioinformatics algorithm was applied to identify leader genes for RCs and PGs. Genes related to RCs and PGs were first identified in PubMed, GenBank, GeneAtlas, and GeneCards databases. The Web-available STRING software (The European Molecular Biology Laboratory [EMBL], Heidelberg, Baden-Württemberg, Germany) was used in order to build the interaction map among the identified genes by a significance score named weighted number of links. Based on the weighted number of links, genes were clustered using k-means. The genes in the highest cluster were considered leader genes. Multilayer perceptron neural network analysis was used as a complementary supplement for gene classification. **Results:** For RCs, the suggested leader genes were *TP53* and *EP300*, whereas PGs were associated with *IL2RG*, *CCL2*, *CCL4*, *CCL5*, *CCR1*, *CCR3*, and *CCR5* genes. **Conclusions:** Our data revealed different gene expression for RCs and PGs, suggesting that not only the inflammatory nature but also other biological processes might differentiate RCs and PGs. *(J Endod 2015;41:877–883)*

Key Words

Endodontics, gene, high throughput biology, TH1, TH2

Inflammatory chronic apical lesions (ICALs) are associated with endodontically involved teeth (1). It has been accepted that the bacterial presence/colonization of the root canal space is the main etiologic factor of ICALs (2, 3). Moreover, host responses such as inflammation (4), angiogenesis (5), and, consequently, bone resorption (1) could modify the severity and prognoses of ICALs. Differential diagnosis among all ICALs is possible only with histopathological examination (6). Moreover, it was suggested that radiographic diagnosis of ICALs should not be used for scientific investigations (7).

Because root canal debridement is the first choice treatment of ICALs (8, 9), most ICALs are impossible to distinguish among radicular cysts (RCs) or periapical granulomas (PGs) (10). Interestingly, there is a discrepancy related to the sample sizes of studies that consider only radiographic examination (8, 11) and studies with tissue sample analyses (2, 4, 5,12–15). The reduction of ICAL specimens for analyses can represent important bias to elucidate ICAL pathology (16).

As in several diseases, ICAL has been related to many forms of controlling gene expression (4, 5,17–19). However, most of these studies focused on isolated genes or metabolic pathways. Bioinformatics has emerged as an important tool for analyses of a plethora of data generated (20–22). Recently, the leader gene approach gave promising results in the context of oral lichen planus (21) and chronic inflammatory periodontitis (20). Genes belonging to the highest rank are defined as leader genes, which are relevant genes in a given cellular process, according to the already available experimental data. The purpose of this study was to characterize the gene expression of RCs and PGs by bioinformatics, interaction network analysis, and neural networks.

Materials and Methods

Bioinformatics and Interaction Network Analysis

The leader gene approach was described previously (20–23). Briefly, key genes involved in RCs or PGs were identified by a search of large-scale gene databases. To determine the primary set of genes, a search considering only human genes was performed on the following databases: PubMed, GenBank, GeneAtlas (22), and GeneCards (24). The gene nomenclature adopted was defined by the Human Genome Organization. Only interactions based on experimental observations described in the public domain and

From the Departments of *Dentistry, †Physiopathology, and ‖Computer Science Universidade Estadual de Montes Claros, Minas Gerais, Brazil; ‡Department of Restorative Dentistry, Faculty of Dentistry, Universidade Federal de Minas Gerais, Minas Gerais, Brazil; and §Department of Restorative Dentistry, Faculty of Dentistry, Universidade de São Paulo, Ribeirão Preto, São Paulo, Brazil.

Address requests for reprints to Dr André Luiz Sena Guimarães, Universidade Estadual de Montes Claros, Hospital Universitário Clemente Faria, Laboratório de Pesquisa em Saúde, 562 Cula Mangabeira Avenue, Santo Expedito, Montes Claros, MG, Brazil 39401-001. E-mail address: andreluizguimaraes@gmail.com

0099-2399/$ - see front matter

http://dx.doi.org/10.1016/j.joen.2015.02.004

available in specific databases were considered with a high degree of confidence (above 0.9; range, 0–0.99) (20–23). With this process, new genes directly linked to RCs and PGs could be identified. Literature data from PubMed, GenBank, GeneAtlas, and GeneCards was performed using a string of pertinent key words chosen by experts; Medical Subject Headings were used to carefully check terms and all their possible Boolean logics–based combinations to avoid false-positive data. After this step, a list of potential "candidate genes" related to RCs or PGs was generated. The initial gene list for RCs and PGs was then expanded using the Web-available software STRING (version 9.05) (The European Molecular Biology Laboratory [EMBL], Heidelberg, Baden-Württemberg, Germany) (25, 26). The STRING software was used to score each interaction to build the interaction map among the identified genes. For every gene identified, we summed combined association scores and adjusted, multiplying to 1000 (20–23), to obtain a single score named weighted number of links (WNL). Based on the WNL, genes were clustered using k-means. Leader genes had the highest rank, and the other genes were termed in a decreased alphabetical way according to their WNL score. Genes with no interactions were defined as orphan genes (21, 23). To evaluate differences among various classes in terms of WNL, analysis of variance (ANOVA) and Tukey-Kramer post hoc tests were used. Statistical significance was set at a P value <.001. Interacting genes were classified as up-regulated, down-regulated, or neutral in respect to RC or PG pathogenesis. Genes that did not exhibit fold expression changes in the disease versus health control condition or genes for which there is no universal consensus in the literature or databases were considered neutral genes. Topologic analysis was performed with Cytoscape (San Diego, CA) (27) and FANMOD (Jena, Thuringia, German) (28), whereas ontologic analysis was performed with Biological Networks Gene Ontology tool (BiNGO) (20–23, 29). Human periodontal ligament cell analyses were also performed and are shown in Supplemental Figure S1 (Supplemental Figure S1 is available online at www.jendodon.com).

Neural Networks

An alternative approach for gene classification inspired by the configuration of the human brain involves the use of neural networks. In this approach, the multilayer perceptron (MLP) neural network was used for gene classification (30–32) as a complementary supplement to the methodology described previously. MLP is 1 specific feedforward network with 1 or more layers between the input and output nodes (hidden layers). Training is achieved of MLP by the backpropagation algorithm (33). Training is achieved of MLP by the back propagation algorithm (30, 33):

1. *Initialization:* Assume that no prior information is available and pick the synaptic weights and thresholds, $\omega_{ji}^{(l)}(n)$, from a uniform distribution whose mean 0 and whose variance is chosen to make the standard deviation of induced local fields of the neurons lie at the transition between the linear and saturated parts of the activation function, φ.
2. *Presentation of training examples:* Present the neural net with an epoch of training examples. For each example in the set $\{(x(n), d(n))\}_{n=1}^{N}$, ordered in some fashion, perform the sequence of forward and backward computations described under points 3 and 4, respectively.
3. *Forward computation:* Let a training example in the epoch be denoted by (x (n), d (n)), with the input vector x (n) applied to the input layer of sensory nodes and the desired response vector d (n) presented to the output layer of computation nodes. Compute the induced local field and function signals of the neural net by proceeding forward through the net, layer by layer. The induced local field $v_j^{(l)}(n)$ for neuron j in layer l is

$$v_j^{(l)}(n) = \sum_{i=0}^{m_0} \omega_{ji}^{(l)}(n) y_i^{(l-1)}(n)$$

where $y_i^{(l-1)}(n)$ is the output (function) signal of neuron i in the previous layer $(l-1)$ at iteration n and $\omega_{ji}^{(l)}(n)$ is the synaptic weight of neuron j in layer l that is fed from neuron i in layer $(l-1)$. For $i = 0$, we have $y_0^{(p-1)}(n) = +1$ and $\omega_{j0}^{(l)}(n) = b_j^{(l)}$ is the bias applied to neuron j in layer l. The output signal of neuron j in layer l is

$$y_j^{(l)} = \varphi_j\left(v_j(n)\right)$$

If neuron j in the first hidden layer (ie, l = 1), set

$$y_j^{(0)}(n) = x_j(n)$$

where $x_j(n)$ is the jth element of the input vector x (n). If neuron j is in the output layer (ie, $l = L$, where L is the number of network layers), set

$$y_j^{(L)} = o_j(n)$$

Compute the error signal

$$e_j(n) = d_j(n) - o_j(n)$$

where $d_j(n)$ is the jth element of the desired output vector d (n).

4. *Backward computation:* Compute the δs (ie, local gradients) of the neural net defined by

$$\delta_j^{(l)}(n) = \begin{bmatrix} e_j^{(L)}(n)\varphi_j'\left(v_j^{(L)}(n)\right) \\ v_j^{(l)}(n)\sum_k \delta_k^{(l+1)}(n)\omega_{kj}^{(l+1)}(n) \end{bmatrix}$$

where the prime in $\varphi_j'(.)$ denotes differentiation with respect to the argument. Adjust the synaptic weights of the neural net in layer l according to the following generalized delta rule (34, 35):

$$\omega_{ji}^{(l)}(n+1) = \omega_{ji}^{(l)}(n) + \alpha\left[\omega_{ji}^{(l)}(n-1)\right] + \eta\delta_j^{(l)}(n) y_j^{(l-1)}(n)$$

where η is the learning rate and α is the momentum constant.

5. *Iteration:* Iterate the forward and backward computations under points 3 and 4 by presenting new epochs of training examples to the neural net until the stopping criterion is met.

Results

Bioinformatics and Interaction Networks

Considering RCs, the preliminary query in GeneCards suggested 25 genes. It was performed as an expansion on STRING, and 10 genes were included (Fig. 1). For RCs, the suggested leader genes were *TP53* and *EP300* (Fig. 2*D*). The difference in WNL scores between the leader genes related to RCs was confirmed by ANOVA with the Tukey post hoc test ($P < .001$). The network for RCs also exhibits a power law