Meta Gene 12 (2017) 72–77

Contents lists available at ScienceDirect

Meta Gene

journal homepage: www.elsevier.com/locate/mgene

An adaptation of particle swarm clustering applied in basal cell carcinoma, squamous cell carcinoma of the skin and actinic keratosis

Fabiano de Oliveira Poswar [a], Laércio Ives Santos [d], Lucyana Conceição Farias [a], Talita Antunes Guimarães [a], Sérgio Henrique Souza Santos [b], Kimberly Marie Jones [c], Alfredo Maurício Batista de Paula [a], Reinaldo Martinez Palhares [e], Marcos Flávio Silveira Vasconcelos D'Angelo [f,*], André Luiz Sena Guimarães [a,*]

[a] Department of Dentistry, State Universit of Montes Claros, Brazil
[b] Institute of Agricultural Sciences, Food Engineering College, Universidade Federal de Minas Gerais (UFMG), Montes Claros, Minas Gerais, Brazil
[c] Center for Research, The Educative Association of Brazil, Montes Claros, Brazil
[d] Campus Montes Claros, Federal Institute of Minas Gerais, Brazil
[e] Department of Electronics Engineering, Federal University of Minas Gerais, Brazil
[f] Department of Computer Science, State Universit of Montes Claros, Brazil

ARTICLE INFO

Article history:
Received 19 October 2016
Revised 16 December 2016
Accepted 20 January 2017
Available online 24 January 2017

Keywords:
Particle swarm clustering
Skin cancers
Potential malignant lesion
Metastasis
Bioinformatics

ABSTRACT

Introduction: This study used the comparison of basal cell carcinoma (BCC), squamous cell carcinoma of the skin (SCC) and actinic keratosis (AK) to test a new method for data set clustering in the leader gene approach.
Methods: Genes related to BCC, SCC and AK, were identified in the databases: OMIM, Genecards and NCBI Gene. A network was built for BCC, SCC and AK using STRING. For each gene, a weighted number of links (WNL) was calculated based on the combined STRING scores. The genes were then clustered according to their WNL and TIS, using an adaptation of particle swarm clustering (PSC) or K-means clustering.
Results: A disagreement between K-means clustering and PSC was observed for both BCC and SCC. PSC suggested completed different leader genes to BCC and SCC. While K-means clustering indicated that *CTNNB1* and *TP53* were associated with BCC and SCC. In contrast, no differences in methods were observed to AK, which had the shorter network. TP53 was the only leader gene for AK.
Conclusion: In conclusion, the current study suggests that PSC is an interesting tool for clustering genes in bioinformatics analyses of prevalent diseases. K-means clustering should be used in the small network. The current study also suggests TP53 may play a central role for AK. Additionally, CTNNB1 seems to be related to BCC, while *CTNNA1* is related to SCC

1. Introduction

Non-melanoma skin cancer, including basal cell carcinoma (BCC) and squamous cell carcinoma of the skin (SCC), are the most common malignancies in humans (Guenther et al., 2015). BCC and SCC share essential characteristics. They both are derived from epidermal keratinocytes and are associated with ultraviolet light exposure, fair skin, and immunosuppression (Madan et al., 2010). Nevertheless, SCC has a greater tendency to metastasize and frequently arises from precursor lesions, such as actinic keratosis (AK) (Chetty et al., 2015), while BCC only very rarely metastasize and it arises directly from healthy skin (Bauer et al., 2011). AK is a lesion related to cumulative sun exposure. After AK is established, it can either evolve to spontaneous remission, remain stable or transform into invasive SCC (Berman and Cockerell, 2013).

The identification of biological markers that could be related to the distinct behaviors of BCC and SCC is an exciting field of research since it has the potential to unveil key elements related to invasion and metastasis (Shimizu et al., 2001; Galer et al., 2011; Yin et al., 2013). Also, the discovery of genes differentially expressed between AK and SCC can be important in understanding the carcinogenesis of keratinocyte-derived neoplasms. In fact, much research has been done comparing these neoplasms, and there is already a good amount of data on this topic (Poswar et al., 2013; de Oliveira Poswar et al., 2015). Bioinformatics has emerged as an important tool for mining relevant databases for significant evidence of specific pathways and, in particular, in identifying genes that exhibit a high level of activity with other genes during lesion-specific disease processes (Poswar Fde et al., 2015). The identification and comparison of disease-specific "leader" genes have been identified as an approach with promising potential for

* Corresponding authors at: Universidade Estadual de Montes Claros, Hospital Universitário Clemente Faria, Laboratório de Pesquisa em Saúde, 562 Cula Mangabeira Avenue, Santo Expedito, 39401-001 Montes Claros, MG, Brazil.
E-mail addresses: marcos.dangelo@unimontes.br (M.F.S.V. D'Angelo), andreluizguimaraes@gmail.com (A.L.S. Guimarães).

http://dx.doi.org/10.1016/j.mgene.2017.01.007
2214-5400/

contributing to new strategies in the prevention and cure of oral cancers that consider the molecular origins and processes of disease (Orlando et al., 2013; Poswar Fde et al., 2015).

Here a new approach for data set clustering, based on an adaptation of PSC, is proposed. The new approach, associated with interaction network analysis, was used to characterize gene expression of BCC, SCC and AK.

2. Materials and methods

2.1. Bioinformatics and systems biology analysis

The leader gene approach has been previously described (Giacomelli and Nicolini, 2006; Covani et al., 2008; Bragazzi et al., 2011; Orlando et al., 2013; Guimaraes et al., 2016a,b). Briefly, key genes involved in BCC, SCC and AK were identified through searches of large-scale gene databases. A search considering only human genes was performed on the following databases: NCBI Gene, OMIM, and Genecards (Rebhan et al., 1997) to determine the primary set of genes. The gene nomenclature of the Human Genome Organization was adopted (HUGO). The keywords used for SCC were "cutaneous squamous cell carcinoma", "squamous cell carcinoma of the skin", "skin squamous cell carcinoma", "cutaneous squamous cell cancer", "skin squamous cell cancer" and "squamous cell cancer of the skin"; for BCC "basal cell carcinoma", "basocellular carcinoma" and "basal cell epithelioma" was used; for AK, "actinic keratosis" was used. After this step, lists of potential "candidate genes" related to BCC, SCC and AK were generated. The initial gene list for BCC, SCC and AK was then expanded using the Web-available software STRING (version 9.05) (von Mering et al., 2005; Jensen et al., 2009). Only interactions based on experimental observations described in the public domain and available in specific databases were considered with a high degree of confidence (above 0.9, range 0–0.99) (Covani et al., 2008; Bragazzi et al., 2011; Orlando et al., 2013). With this process, new genes directly linked to BCC, SCC and AK were identified. Literature-based data from PubMed were used to avoid false positives in the analysis of data. The STRING software was used to score each interaction and build maps among the identified genes. For every gene identified, we summed combined association scores specifically in the network and adjusted by multiplying to 1000 (Giacomelli and Nicolini, 2006; Covani et al., 2008; Bragazzi et al., 2011; Orlando et al., 2013), to obtain a single score, named a weighted number of links (WNL). The total interaction score (TIS) represents all gene interactions in the entire STRING database. To obtained TIS value, all interactions of a gene in the whole STRING database were summed and adjusted by multiplying to 1000. Genes with no interactions were defined as orphan genes (Bragazzi et al., 2011; Orlando et al., 2013).

The topological analysis was carried out with Cytoscape (Shannon et al., 2003) and FANMOD (Wernicke and Rasche, 2006), while ontological analysis was performed with BinGO (Maere et al., 2005; Giacomelli and Nicolini, 2006; Covani et al., 2008; Bragazzi et al., 2011; Orlando et al., 2013).

2.2. Clustering procedures

Based on the WNL and TIS (Bragazzi et al., 2011; Guimaraes et al., 2016a,b), genes were clustered, using K-Means Clustering or IPC. Leader genes are identified as genes with the highest rank of associations; other genes were termed by decreasing alphabetical identifiers according to their WNL and TIS score. The category with higher WNL and lower TIS is the more important and was considered as category A or leader genes. All other categories apart from leader genes were not taken into account for analyses purpose.

After K-Means Clustering, ANOVA, and Tukey-Kramer post hoc tests were applied to certify the results. Statistical significance was set at a P value of <0.001 as performed before (Covani et al., 2008; Poswar Fde et al., 2015; Guimaraes et al., 2016a,b; Sobrinho-Santos et al., 2016)

The particle swarm clustering (PSC) algorithm was proposed in to be a grouping tool inspired by social human behavior (Eberhart et al., 2001). This algorithm is based on Particle Swarm Optimization (PSO). In PSO was added in the auto-organization term (Kohonen, 1990) in particles process move.

In PSC the particles are moved to become prototypes of groups in algorithm execution. Therefore, each input pattern is presented to the particle swarm, and the most similar particle is moved in the direction of the entry pattern, influenced by inertia, social and cooperation terms. The particle velocity is updated by Eq. (1).

$$V_v(t+1) = \omega V_v(t) + \varphi_1 \otimes (Pbest_{vj}(t) - X_v(t)) + \varphi_2 \otimes (Gbest_j(t) - X_v(t)) + \varphi_3 \otimes (Y_j - X_v(t)) \quad (1)$$

where ω is the inertia term used to prevent particles with very high speed from moving out of the universe of discourse, the terms φ_1, $\varphi_2$and φ_3 are vectors of random weights, generated by an uniform distribution on the interval [0,1], used to weight, respectively, the terms of memory, cooperation and, self-organization of each particle, $\otimes$ represents a multiplication point to point operator, $pbest_i$ is the *i-th* particle best position (position of greater similarity with input pattern) and *Gbest* represents the best position of all particles in relation to the input pattern. Each particle position is updated by Eq. (2).

$$X_v(t+1) = X_v(t) + V_v(t+1) \quad (2)$$

Each particle moves through search space considering only its similarity with the input pattern, and it does not use cost or quality solution information for moving. Thus, the PSC is characterized as a not supervised algorithm. The following steps of PSC are performed in each iteration:

1. Take the input patterns;
2. Start k particles randomly;
3. For each pattern, determine the particle with greater similarity (winning particle);
4. Upgrade the terms *Pbest* and *Gbest* of the winning particle, if necessary;
5. Update winning particle velocity according to (1) and its position according to (2);
6. Decrease the value of ω for some linear function;
7. Repeat step 3 to step 6 until some stopping criterion is satisfied.

If no winning particle in the algorithm execution, its speed is updated in the direction of the most successful particle.

To enable the PSC automatically find the number of groups, two changes were made in the PSC algorithm: growth operator (3) and the pruning particles. In the algorithm execution, the concentration level of each particle is determined. If a particle has its level of concentration >0 becomes a candidate to be cloned. If the similarity between the particle candidate and the most similar input pattern is greater than a threshold (ε), the particle will be cloned. The particle clone is positioned between the cloned particle and the input pattern most similar. The pruning operator removes particles at concentration levels equal to zero. PSC algorithm was included in the Supplementary material.

3. Results

3.1. Network features

A total of 382 distinct genes related to BCC were identified, precisely 372 in the Genecards database, 121 in the NCBI gene database and 11 in the OMIM database (Fig. 1A and Supplementary material). The network for BCC exhibits a power-law behavior (correlation: 0.713; R2:0. 679), in agreement with the scale-free theory of networks (Supplementary material), as well as more feed-forward loops than would be expected in a random network (11.58%; Z-score: 33.39). SCC was related to 178