Coordinated Markov Modeling of Cancer Metastasis from Multiple Primary Sites

Hyunggu Jung[1,*], Anthony Law[2], Esther Wu[1], Mark E. Whipple[1,2]

[1]*Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, WA, U.S.A.*
[2]*Department of Otolaryngology-Head and Neck Surgery, University of Washington, Seattle, WA, U.S.A.*
*hyunggu@uw.edu

ABSTRACT

Objectives: Planning effective treatment of cancer of the head and neck requires predicting the probability of microscopic tumor spread to regional lymph nodes. Previous studies demonstrated that Markov chain models are feasible methods to predict tumor spread in the lymphatic system from individual primary tumor sites in the head and neck. However, little is still known about how to utilize data from multiple primary sites with overlapping lymphatic drainage to improve model performance. In this study, we investigated whether a Markov chain model that was based upon lymphatic drainage pathways could predict the probability of metastasis to individual nodal groups. Further, we tested if a Markov model that uses parameters obtained by training with data from two different primary sites performed better than a model trained with data from a single primary site.

Methods: We created a two-dimensional Markov chain model where the row of the model represents the metastatic progression along the lymphatic pathway and the column indicates the T-stages for the primary tumor location. To estimate the parameters of the model, we used two data sets as training sets: one from the records of 50 patients with non-treated, non-recurrent squamous cell carcinoma (SCCA) of the oral tongue and 10 patients with SCCA of the buccal mucosa, which presented to the University of Washington (UW) head and neck tumor board over a 3.5-year period. We ran the model with all possible parameters with a step size of 0.1 to identify the upper and lower bound of the parameters that fit with the training sets. We then determined the parameters when the output of the model was closest to the training sets after running the model again for any parameters with a step size of 0.05 within the range. For validating the model, we compared the closeness (i.e., cosine similarity) between the output of the model and a test set derived from the Cancer Genome Atlas (TCGA), which included 81 patients with SCCA of the oral tongue and 9 patients with SCCA of the buccal mucosa. For each of the two primary sites, we compared the performance between the Markov model trained with data from the primary site only to the performance of the model that was trained using data from both primary sites. For comparison, we also measured the cosine similarity between the training and test sets.

Results: The cosine similarity between the output of the Markov chain model and the test set was greater than the cosine similarity between the training and test sets (see Table 1). Also, the output of the model using the parameters trained by combined data sets from two different primary sites (i.e., buccal mucosa and oral tongue) showed better prediction than models trained using a single data set.

Conclusion: We validated the Markov chain model for predicting tumor spread using the TCGA data with the estimated parameters from different primary sites. The results of our study indicated that our Markov chain model may accomplish better performance when utilizing multiple data sets from different primary sites than a model trained with data from a single primary site.

KEYWORDS
Markov chain, Oral cancer, Metastasis

Case for buccal mucosa	Cosine similarity with TCGA	Case for oral tongue	Cosine similarity with TCGA
Buccal mucosa data from UW	0.279	Oral tongue data from UW	0.889
Model output trained with buccal mucosa data	0.584	Model output trained with oral tongue data	0.914
Model output trained with combined data	0.587	Model output trained with combined data	0.967

Table 1. Cosine similarity between two data sets: 1) UW patient data and TCGA data, 2) model output trained with a single data set and TCGA data, 3) model output trained with combined data and TCGA data