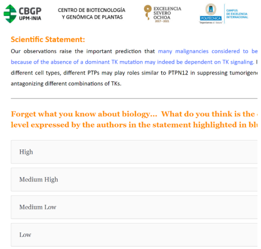
Feb 15, 2019

Scholarly Certainty Survey and Analysis

DOI
dx.doi.org/10.17504/protocols.io.x6nfrde



CBGP UPM-INIA
CENTRO DE BIOTECNOLOGÍA Y GENÓMICA DE PLANTAS
EXCELENCIA SEVERO OCHOA 2017 - 2021
POLITÉCNICA
CAMPUS DE EXCELENCIA INTERNACIONAL

Scientific Statement:

Our observations raise the important prediction that many malignancies considered to be
because of the absence of a dominant TK mutation may indeed be dependent on TK signaling. I
different cell types, different PTPs may play roles similar to PTPN12 in suppressing tumorigen
antagonizing different combinations of TKs.

Forget what you know about biology... What do you think is the
level expressed by the authors in the statement highlighted in bl

High

Medium High

Medium Low

Low

Mark D D. Wilkinson[1], Mario Prieto Godoy[2]

[1]CBGP UPM-INIA; [2]Centro de Biotecnología y Genómica de Plantas UPM – INIA

Mark D D. Wilkinson
Centre for Plant Biotechnology and Genomics UPM – INIA

DOI: dx.doi.org/10.17504/protocols.io.x6nfrde

Protocol Citation: Mark D D. Wilkinson, Mario Prieto Godoy 2019. Scholarly Certainty Survey and Analysis. **protocols.io** **https://dx.doi.org/10.17504/protocols.io.x6nfrde**

Protocol status: In development
We are still developing and optimizing this protocol; however, it has been used to generate results that are under peer-review. Future edits to this protocol will take reviewer comments into account.

Created: February 15, 2019

Last Modified: February 15, 2019

Protocol Integer ID: 20398

Keywords: questionnaire, certainty survey

Abstract

Using the TAC Biomedical Summarization Corpus **(Min-Yen)**, we extracted 45 manually curated scholarly assertions (selection described below). Based on these, we generated three independently-executed Web-based questionnaires where researchers in the biomedical domain, from different linguistic groups but in comparable research institutes, were presented with assertions and asked to categorize the strength of those assertion into 4 (High, Medium High, Medium Low and Low), 3 (Category 1, Category 2 and Category 3) or 2 (Relatively High and Relatively Low) certainty levels over the three questionnaires. G Index **(Holley & Guilford, 1964)** coefficient analysis was applied to determine the degree of agreement between annotators. We then extracted the essential features of inter-rater agreement from the questionnaire data using Principal Component Analysis (PCA). Afterward, we categorized our collection of statements in clusters using k-means algorithm **(Jolliffe, 2011)** **(Dunham, 2006)**. Finally, an automated classifier model was generated using deep-learning techniques over the results of this study (manuscript under preparation). This was used to construct exemplar scholarly assertions capturing certainty metadata, and published as machine-readable NanoPublication.

Materials

The 45 text blocks used in the three surveys were extracted from published articles related to genetic and molecular issues, and were selected from the “Citation Text” and “Reference text” portions of the TAC 2014 Biomedical Summarization Track. Each text block contained a sentence or sentence fragment representing a single scholarly assertion that we asked the respondents to evaluate, highlighted in blue, with the remainder of the text being provided for additional context. These assertions were selected using different epistemic modifiers, such as modal verbs, qualifying adverbs and adjectives, references and reporting verbs **(De Waard & Maat, 2012)**An example survey interface presentation is shown in the image.
Participants of the surveys were engaged using the platform Survey Gizmo **(“Survey Gizmo”)** (S1), and Qualtrics **(“Qualtrics”)** (S2 & S3)- two online platforms dedicated to Web-based questionnaires. The change in survey platform was based only on cost and availability; the two platforms have largely comparable interfaces with respect to data-gathering fields such as response-selection buttons and one-question-per-page presentation, with the primary differences between the platforms being aesthetic (color, font, branding).

Survey Design

1 We designed three surveys - S1, S2 and S3 - where respondents were asked to assign certainty based on a number of certainty categories - 4, 2, and 3 respectively for surveys S1, S2, and S3. The surveys used identical corpori of biomedically-oriented scholarly statements. To minimize the bias of prior exposure to the corpori, the surveys were deployed over three comparable but distinct groups of researchers, all of whom will have sufficient biomedical expertise to understand the statements in the corpus.

All participants were presented assertions selected randomly from the 45 in the corpus - 15 assertions in S1, which we raised to 20 assertions in S2 and S3 in order to obtain deeper coverage of the statement set. In S1, participants had to assess the certainty of every highlighted sentence fragment based on a 4-point scale with the following response options: High, Medium High, Medium Low, and Low. A 2-point scale was used for S2: Relatively High and Relatively Low and 3-point numerical scale for S3: 1, 2 or 3. In addition to the assessment of certainty, for each sentence fragment assertion, subjects were asked to indicate their impression of the basis of the assertion, on a single-answer question, with the options: Direct Evidence, Indirect Evidence/Reasoning, Speculation, Citation and I don't know.

Survey Distribution and Participant Selection

2 Participation in the surveys was primarily achieved through personal contact with department leads/heads of 5 institutions with a focus on biomedical/biotechnology research. For S1, the major set of participants came from the Centro de Biotecnologia y Genomica de Plantas (UPM-INIA), Spain. It was conducted between November and December of 2016. S2 was executed by members from the Leiden University Medical Center, Netherlands, between November and December of 2017. S3 was conducting between October and November of 2018 by members from University Medical Center Utrecht, Cell Press and the Agronomical Faculty of Universidad Politécnica of Madrid. Participation was anonymous and no demographic data was collected.

Statistical analysis

3 We evaluate each survey by quantifying the degree of agreement between participants who were presented the same assertion, with respect to the level of certainty they indicated was expressed in that statement given the provided categories for that survey.

Agreement between participants was assessed by Holley and Guilford's G Index of agreement **(Holley & Guilford, 1964)** which is a variant of Cohen's Weighted Kappa **(Cohen, 1968)**Ideally G measures the agreement between participants . It was performed based on the following formula:

$$G = ProbabilityObserved(Po) - ProbabilityByChance(Pc)/1 - Pc$$

The key difference between Kw and G is in how chance agreement (Pc) is estimated. According to **(Xu & Lorber, 2014)** "G appears to have the most balanced profile, leading us to endorse its use as an index of overall interrater agreement in clinical research". G is defined a priori, being homogeneously distributed among categories as the inverse of the number of response categories **(Xu & Lorber, 2014)**thus making G=0.25 for S1; G=0.50 for S2; and G= 0.33 for S3. The accepted threshold for measuring agreement and its interpretation has been suggested by Landis & Koch, 1977 **(Landis, Richard Landis & Koch, 1977; Viera & Garrett, 2005)**aas follows: 0.2 = Poor, 0.21 - 0.4 = Fair, 0.41 - 0.60 = Moderate, 0.61 - 0.80 = Substantial, 0.81 - 1.00 = Almost Perfect. Anything other than the 'Poor' category is considered by other studies to have achieved an acceptable level of agreement. **(Deery et al., 2000)** **(Lix et al., 2008)**

We then investigated the ideal number of clusters in which statements group in accordance with their certainty levels. To estimate this, Hierarchical Clustering analysis (HCA) and the Spearman correlation test were performed to determine certainty category-association between questionnaires, using the shared classified statements in that category as the metric. All Spearman interactions are based on hypothesis testing. To determine the importance of the results, p-values were generated as an indicator of the existence of correlation between certainty categories.

Finally, we applied Principal Component Analysis (PCA) to the result-sets, and utilized K-means to identify cluster-patterns within the PCA data, which would reflect groups of similar "human behaviors" to individual questions under all three survey conditions. Finally, to determine the optimal K (cohesion of the clusters), several indices were analysed using the R package NbClust **(Charrad et al., 2014)**NbClust provides 30 different indices for determining the best number of clusters based on the majority rule, such as Gap statistic or Silhouette method. Membership in these clusters was evaluated via Jaccard similarity index comparing, pairwise, all three clusters from each of the three surveys to determine which clusters were most alike.