Bioinformatics, 35(24), 2019, 5365–5366
doi: 10.1093/bioinformatics/btz520
Advance Access Publication Date: 2 July 2019
Applications Note

Data and text mining

CytoGPS: a web-enabled karyotype analysis tool for cytogenetics

Zachary B. Abrams[1], Lin Zhang[2], Lynne V. Abruzzo[3], Nyla A. Heerema[3], Suli Li[1], Tom Dillon[2], Ricky Rodriguez[2], Kevin R. Coombes[1] and Philip R. O. Payne[2,]*

[1]Department of Biomedical Informatics, Wexner Medical Center, The Ohio State University, Columbus, OH 43210, USA, [2]Institute for Informatics, Washington University School of Medicine in St. Louis, St. Louis, MO 63108, USA and [3]Department of Pathology, Wexner Medical Center, The Ohio State University, Columbus, OH 43210, USA

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on March 12, 2019; revised on June 12, 2019; editorial decision on June 16, 2019; accepted on June 28, 2019

Abstract

Summary: Karyotype data are the most common form of genetic data that is regularly used clinically. They are collected as part of the standard of care in many diseases, particularly in pediatric and cancer medicine contexts. Karyotypes are represented in a unique text-based format, with a syntax defined by the International System for human Cytogenetic Nomenclature (ISCN). While human-readable, ISCN is not intrinsically machine-readable. This limitation has prevented the full use of complex karyotype data in discovery science use cases. To enhance the utility and value of karyotype data, we developed a tool named CytoGPS. CytoGPS first parses ISCN karyotypes into a machine-readable format. It then converts the ISCN karyotype into a binary Loss-Gain-Fusion (LGF) model, which represents all cytogenetic abnormalities as combinations of loss, gain, or fusion events, in a format that is analyzable using modern computational methods. Such data is then made available for comprehensive 'downstream' analyses that previously were not feasible.
Availability and implementation: Freely available at http://cytogps.org.
Contact: prpayne@wustl.edu

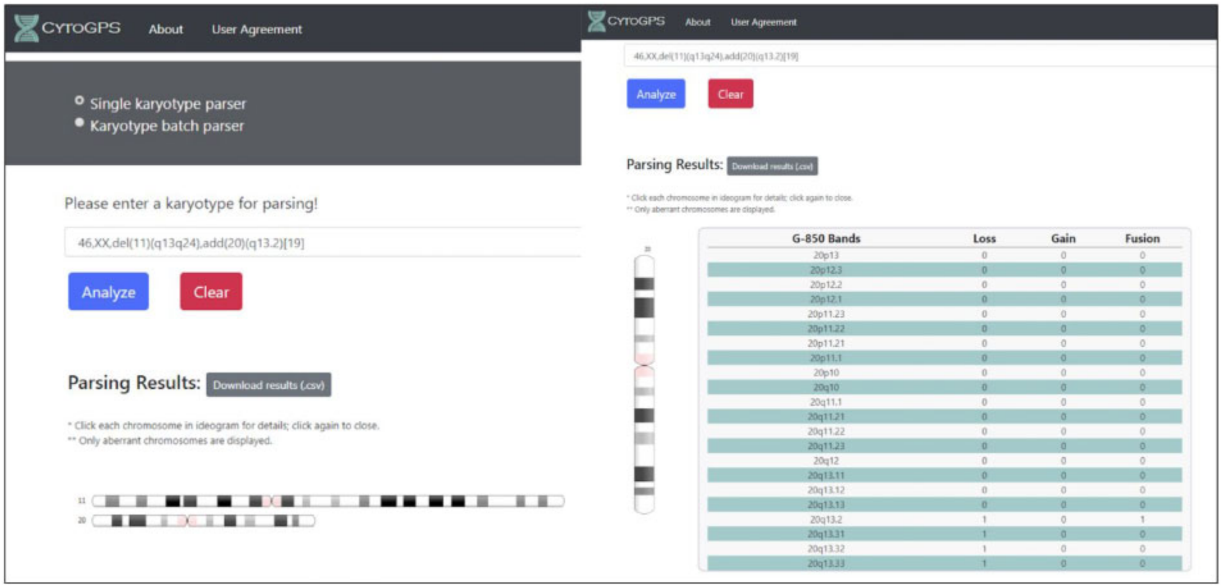
1 Introduction

Karyotype data are the most common form of genetic data regularly employed in clinical medicine. They are stored in a text-based standard format with a syntax defined by the International System for human Cytogenetic Nomenclature or ISCN (McGowan-Jordan *et al.*, 2016). This international standard is updated regularly and dictates how karyotypes should be written and stored. Although trained cytogeneticists are able to read ISCN, the nomenclature is formatted as a semi-context-free and semi-structured language, thus making it incompatible with comprehensive computational analysis. This limitation explains why past attempts to parse karyotype data have only been marginally successful.

To make karyotype more readily available for computational analyses, we developed a rules-based grammar parser called CytoGPS. The parser identifies the relevant cytogenetic information in an ISCN-encoded karyotype, and then uses a mapping system to translate that information into a computer readable format, which we refer to as the Loss-Gain-Fusion (LGF) model (Abrams *et al.*, 2015). In this note, we describe a website (http://cytogps.org) that we built for individual researchers so that they can quickly and easily use CytoGPS. The website provides both single karyotype and batch karyotype processing. It enables researchers to automatically process thousands of karyotypes written in the short form ISCN nomenclature. Further, CytoGPS supports the development of future, structured databases based on the LGF model, which we believe will further cytogenetic research by unlocking the potential currently locked within non-computable ISCN-encoded karyotypes.

2 Implementation

The foundation of CytoGPS is a grammar-based parser that was created using Antlr, a system designed to construct grammar sets



○ Single karyotype parser
● Karyotype batch parser

Please enter a karyotype for parsing!

46,XX,del(11)(q13q24),add(20)(q13.2)[19]

Parsing Results:

* Click each chromosome in ideogram for details; click again to close.
** Only aberrant chromosomes are displayed.

G-850 Bands	Loss	Gain	Fusion
20p13	0	0	0
20p12.3	0	0	0
20p12.2	0	0	0
20p12.1	0	0	0
20p11.23	0	0	0
20p11.22	0	0	0
20p11.21	0	0	0
20p11.1	0	0	0
20p10	0	0	0
20q10	0	0	0
20q11.1	0	0	0
20q11.21	0	0	0
20q11.22	0	0	0
20q11.23	0	0	0
20q12	0	0	0
20q13.11	0	0	0
20q13.12	0	0	0
20q13.13	0	0	0
20q13.2	1	0	1
20q13.31	1	0	0
20q13.32	1	0	0
20q13.33	1	0	0

Fig. 1. CytoGPS website. This shows the karyotype analysis landing page, which describes general information about how to use the web based analysis tool. The left panel shows how results are displayed after a karyotype has been parsed. The user is shown which chromosomes have cytogenetic abnormalities and is provided with a downloadable file of the results. When a user clicks on a chromosome, they see a table in the right panel that shows each cytogenetic region on that particular chromosome and how those regions were affected

(Parr, 2013). Using this parser, CytoGPS can translate ISCN-encoded karyotypes into a parse tree that is capable of decomposing a given karyotype into subcomponents. The resulting parse tree can then be traversed to extract relevant information from different branches and leaves of the tree, thus extracting biological information from the ISCN karyotype.

When parsing a karyotype, the most important information that is encoded is the description of cytogenetic abnormalities located on specific chromosome bands. This information can be thought of as the event and its location. These event/location pairs are translated by CytoGP using a rule-based mapper that contains an enumeration of all ISCN cytogenetic abnormalities as well as a biological interpretation of those abnormalities in the context of losses, gains and fusions. This rules-based system ultimately translates the event/location information from the ISCN karyotype into a binary model that represents each such Loss-Gain-Fusion event (what we call the LGF model). In the LGF model, each cytogenetic band is represented three times—once each for loss, gain and fusion, preserving the cytogenetic meaning of the karyotype.

For ease of use, we developed a website to allow researchers and clinicians the ability to access CytoGPS in an easy and highly usable manner (Fig. 1). To facilitate access to the tool, we made a few key design decisions. First, the homepage gives an overview of CytoGPS functionality. Users may then navigate to an interactive set of interface components, where more detailed descriptions about required inputs to the tools are provided. Third, we include detailed examples and example datasets in both the single karyotype analysis and batch karyotype analysis so that users can explore the system to determine whether CytoGPS can help them meet their research needs. All of these components provide clear visual feedback to the user on the options they have selected as well as how the single karyotype analysis results can be shown. Although we only provide data visualization for individual karyotype analysis, users can individually assess each affected chromosome within their input karyotype via an ideogram that presents each chromosome in horizontal view. By clicking on an individual chromosome, users can access a vertical view of that chromosome that displays more detailed information about each cytogenetic band (Fig. 1). This visually demonstrates how the CytoGPS algorithm translates the ISCN data into the LGF model. For batch karyotypes, we return a file containing the results for each individual karyotype in the batch along with statistics summarizing the entire cohort of karyotypes. These statistics include the percentage of the input population that had specific cytogenetic abnormalities, so users can easily determine the defining cytogenetic events for their dataset.

3 Conclusion

CytoGPS is a unique research tool for the computational analysis of karyotype data. By developing a free, user-friendly website, we hope to improve the dissemination and usability of the CytoGPS algorithm to aid a wider number of research groups in analyzing their cytogenetic data.

Funding

This work was supported by the National Library of Medicine (NLM) grant number T15 LM011270, the National Cancer Institute grant number R03 CA235101 and by Pelotonia Intramural Research Funds from the James Cancer Center.

Conflict of Interest: none declared.

References

Abrams,Z.B. *et al.* (2015) Text mining and data modeling of karyotypes to aid in drug repurposing efforts. *Stud. Health Technol. Inform.*, **216**, 1037.

McGowan-Jordan,J. *et al.* (2016) *ISCN 2016: An International System for Human Cytogenomic Nomenclature*. Karger Medical and Scientific Publishers, Basel.

Parr,T. (2013) *The Definitive ANTLR 4 Reference*. Pragmatic Bookshelf, Dallas, TX.