Bioinformatics, 32(12), 2016, 1888–1890
doi: 10.1093/bioinformatics/btw084
Advance Access Publication Date: 15 February 2016
Applications Note

Structural bioinformatics

BALL-SNPgp—from genetic variants toward computational diagnostics

Sabine C. Mueller[1,2,*], Christina Backes[1], Alexander Gress[3], Nina Baumgarten[1], Olga V. Kalinina[3], Andreas Moll[1], Oliver Kohlbacher[4,5,6], Eckart Meese[2] and Andreas Keller[1]

[1]Chair for Clinical Bioinformatics, Saarland University, Saarbrücken 66123, Germany, [2]Department of Human Genetics, Saarland University, Homburg 66421, Germany, [3]Max Planck Institute for Informatics, Saarland University, Saarbrücken 66123, Germany, [4]Applied Bioinformatics, Center for Bioinformatics, Quantitative Biology Center, [5]Department of Computer Science, University of Tuebingen, Tübingen 72076, Germany, and [6]Biomolecular Interactions, Max Planck Institute for Developmental Biology, Tübingen 72076, Germany

*To whom correspondence should be addressed.

Associate Editor: Anna Tramontano

Received on October 11, 2015; revised on January 12, 2016; accepted on February 7, 2016

Abstract

Summary: In medical research, it is crucial to understand the functional consequences of genetic alterations, for example, non-synonymous single nucleotide variants (nsSNVs). NsSNVs are known to be causative for several human diseases. However, the genetic basis of complex disorders such as diabetes or cancer comprises multiple factors. Methods to analyze putative synergetic effects of multiple such factors, however, are limited. Here, we concentrate on nsSNVs and present BALL-SNPgp, a tool for structural and functional characterization of nsSNVs, which is aimed to improve pathogenicity assessment in computational diagnostics. Based on annotated SNV data, BALL-SNPgp creates a three-dimensional visualization of the encoded protein, collects available information from different resources concerning disease relevance and other functional annotations, performs cluster analysis, predicts putative binding pockets and provides data on known interaction sites.

Availability and implementation: BALL-SNPgp is based on the comprehensive C++ framework Biochemical Algorithms Library (BALL) and its visualization front-end BALLView. Our tool is available at www.ccb.uni-saarland.de/BALL-SNPgp.

Contact: ballsnp@milaman.cs.uni-saarland.de

1 Introduction

Non-synonymous single nucleotide variants (nsSNVs) play a crucial role for understanding the genetic basis of many diseases, including diabetes and cancer. Over 85% of known nsSNVs are currently associated with a specific disease. While the advent of next-generation sequencing (NGS) increases the amount of identified nsSNVs, the experimental analysis of their pathogenic influence is laborious and time-consuming (Thusberg *et al.*, 2011). Consequently, several approaches to predict the functional impact of nsSNVs *in silico* have been developed. In a previous study, we evaluated the prediction concordance and congruency of 13 state-of-the-art pathogenicity prediction tools (Mueller *et al.*, 2015a). To this end, we were able to address fundamental problems of current methods centering at the ‘one SNV, one phenotype’ paradigm. A human individual inherits several nsSNVs simultaneously. These nsSNV sets may exhibit synergistic effects and, thus, contribute to pathogenic phenotypes. Besides browser-based tools such as MuPIT Interactive (Niknafs *et al.*, 2013) providing three-dimensional (3D) visualizations, computational methods to assess the impact of nsSNV sets on a protein are missing.

In the first approach to analyze nsSNV sets, we developed the proof of concept tool BALL-SNP (Mueller *et al.*, 2015b) that visualizes mutated residues within 3D protein structures and provides available database annotations on clinical significance and disease association for them. Here we present BALL-SNPgp, which expands our previous tool to improve the assessment of pathogenic relevance in clinical diagnostics, especially when no 3D structural information is available. Based on NGS output that relies on the Variant Call Format and includes gene-based annotation, BALL-SNPgp searches for a 3D structure or a homologous 3D model and pathogenicity information in available databases, predicts pathogenicity, protein stability changes and active sites, performs cluster analyses on mutated residues and assigns protein interaction sites as annotated in PiSITE (Higurashi *et al.*, 2009) on the residue level.

2 Features and methods

BALL-SNPgp allows two inputs: a standard SNV-annotation format from ANNOVAR (Wang *et al.*, 2010) and a simple tab-separated file including gene-based annotation (examples on the Web site). Based on the gene identifier, BALL-SNPgp searches for all available 3D structures in the Protein Data Bank (PDB). If no PDB structure is available, we search for available 3D models in ModBase (Pieper *et al.*, 2004). Given a 3D structure, BALL-SNPgp visualizes the encoded protein with the amino acid substitutions introduced by the given nsSNVs and extracts available information on pathogenicity from the UniProtKB, dbSNP and ClinVar. In addition, DrugBank and PiSITE are queried for information whether the protein in question is a known drug target and whether it contains already detected interaction sites, respectively. To study putative cumulative effects, BALL-SNPgp furthermore performs a hierarchical bottom-up cluster analysis on the amino acid residues affected by the nsSNV. The proximity to protein active sites may also provide crucial insights into pathogenic effects. If the active site is not known, BALL-SNPgp predicts it using the Putative Active Sites with Spheres method (Caetano Traina, 2000). Figure 1 illustrates the overall BALL-SNPgp workflow, which is automatically processed given an input file.

All generated information is visualized in the molecular 3D view, where particular properties additionally can be highlighted. An HTML-based information page provides an intuitive instrument for interaction with the data collected from different sources to enable systematic analysis of the results.

2.1 Implementation details

BALL-SNPgp is based on the proof-of-concept implementation of BALL-SNP, which is built on the comprehensive C++ framework Biochemical Algorithms Library (BALL) and its visualization front-end BALLView (Hildebrandt *et al.*, 2010). Because BALL-SNPgp highly relies on 3D information, it allows the use of both, available 3D structures and homologous 3D models. A valid 3D model from ModBase is restricted to a minimum sequence identity of 60% and comprises at least one of the provided amino acid substitutions. However, the result of BALL-SNPgp depends on the quality of the available 3D model. The pathogenicity prediction is performed by the state-of-the-art methods PolyPhen2, PhD-SNP, PANTHER and PROVEAN. I-Mutant2.0 is applied to predict protein stability changes introduced by amino acid substitutions. The 3D model search and the predictions by third-party tools are performed via a constructed compute server to ensure straight-forward maintenance and simple installation of BALL-SNPgp. The used databases are integrated within the source code to allow user-friendly queries. The generated information is presented on a QtWebKit-based HTML widget including standard HTML hyperlinks to interact with the molecular 3D view.

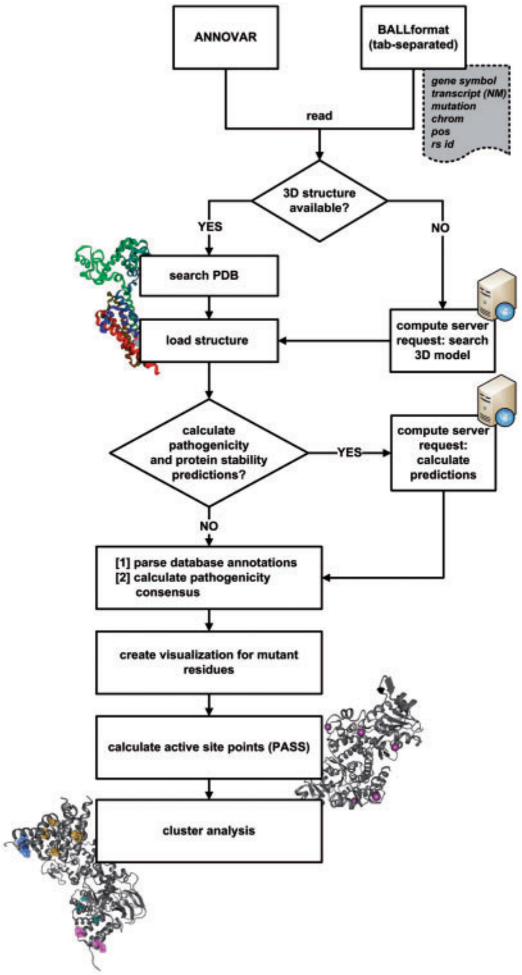
ANNOVAR
BALLformat (tab-separated)
gene symbol
transcript (NM)
mutation
chrom
pos
rs id
read
3D structure available?
YES
NO
search PDB
load structure
compute server request: search 3D model
calculate pathogenicity and protein stability predictions?
YES
compute server request: calculate predictions
NO
[1] parse database annotations
[2] calculate pathogenicity consensus
create visualization for mutant residues
calculate active site points (PASS)
cluster analysis

Fig. 1. BALL-SNPgp workflow given annotated NGS data

3 Conclusion

In this work, we present BALL-SNPgp, a software tool to analyze and identify nsSNV candidates for computational diagnostics. It is based on a standard molecular modeling framework, allows the use of standard NGS output, embeds fundamental nsSNV information and performs various analyses of the 3D structure of the mutated protein to gain insights into putative cumulative effects of nsSNVs. The well-defined user interface and the clear presentation of the generated information are appealing to non-expert users, as well as bioinformatics professionals. In the future, BALL-SNPgp will be extended to include the possibility to analyze binding of the affected protein to its interaction partners and small molecules to facilitate drug design studies.

Funding

This work has been supported by the Best Ageing grant 306031 from the European Union.

Conflict of Interest: none declared.

References

Caetano Traina,A.T.Jr, *et al.* (2000) Fast feature selection using fractal dimensions. *Proceedings of the 15th Brazilian Symposium on Databases* XV Simpósio Brasileiro de Banco de Dados, Joao Pessoa, Paraíba, Brasil, Anais.

Higurashi,M. *et al.* (2009) PiSite: a database of protein interaction sites using multiple binding states in the PDB. *Nucleic Acids Res.*, **37**, D360–D364.

Hildebrandt,A. *et al.* (2010) BALL–biochemical algorithms library 1.3. *BMC Bioinformatics*, **11**, 531.

Mueller,S.C. *et al.* (2015a) Pathogenicity prediction of non-synonymous single nucleotide variants in dilated cardiomyopathy. *Brief. Bioinform.*, **16**, 769–779.

Mueller,S.C. *et al.* (2015b) BALL-SNP: combining genetic and structural information to identify candidate non-synonymous single nucleotide polymorphisms. *Genome Med.*, **7**, 65.

Niknafs,N. *et al.* (2013) MuPIT interactive: webserver for mapping variant positions to annotated, interactive 3D structures. *Hum. Genet.*, **132**, 1235–1243.

Pieper,U. *et al.* (2004) MODBASE, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Res.*, **32**, D217–D222.

Thusberg,J. *et al.* (2011) Performance of mutation pathogenicity prediction methods on missense variants. *Hum. Mutat.*, **32**, 358–368.

Wang,K. *et al.* (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.*, **38**, e164.