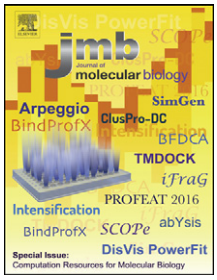

Intensification: A Resource for Amplifying Population-Genetic Signals with Protein Repeats

Jieming Chen [1,2,3], Bo Wang [5], Lynne Regan [1,2,3,5,†] and Mark Gerstein [1,2,3,4,†]

*1 - **Program in Computational Biology and Bioinformatics,** Yale University, New Haven, CT 06520, USA*
*2 - **Integrated Graduate Program in Physical and Engineering Biology,** Yale University, New Haven, CT 06520, USA*
*3 - **Department of Molecular Biophysics and Biochemistry,** Yale University, New Haven, CT 06520, USA*
*4 - **Department of Computer Science,** Yale University, New Haven, CT 06520, USA*
*5 - **Department of Chemistry,** Yale University, New Haven, CT 06520, USA*

***Correspondence to Mark Gerstein:** Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, USA. mark@gersteinlab.org*
http://dx.doi.org/10.1016/j.jmb.2016.12.003
Edited by Michael Sternberg

Abstract

Large-scale genome sequencing holds great promise for the interpretation of protein structures through the discovery of many, rare functional variants in the human population. However, because protein-coding regions are under high selective constraints, these variants occur at low frequencies, such that there is often insufficient statistics for downstream calculations. To address this problem, we develop the Intensification approach, which uses the modular structure of repeat protein domains to amplify signals of selection from population genetics and traditional interspecies conservation. In particular, we are able to aggregate variants at the codon level to identify important positions in repeat domains that show strong conservation signals. This allows us to compare conservation over different evolutionary timescales. It also enables us to visualize population-genetic measures on protein structures. We make available the Intensification results as an online resource (http://intensification.gersteinlab.org) and illustrate the approach through a case study on the tetratricopeptide repeat.

Introduction

The combined efforts from large-scale human sequencing projects and clinical sequencing have given rise to an exponentially increasing number of human sequences in the recent years [1–3]. With substantial drop in the sequencing cost and improvement in sequencing technologies and data processing capabilities, we now have the ability to generate a huge catalog of variants that exist in the human population in a fairly rapid and high-throughput fashion. One of the challenges is to provide functional annotations for these variants efficiently and accurately.

Much of the variant annotation work has been performed in the protein-coding regions. A non-synonymous mutation is considered functionally disruptive if it occurs in a region of high conservation, which is considered to be important evolutionarily [4]. Evolutionary conservation can be observed at different levels. Interspecies comparison can pick out fixed differences between the dominant homologous sequences of the chosen species across their phylogeny over a long evolutionary time [5–7]. At a more recent timescale, intraspecies conservation (across a population) has been observed over specific sites in a few large-scale sequencing studies by aggregating variants over a region or site within the human population [2,8,9]. However, all protein-coding regions are, in general, under high selection pressure. As such, almost all positions in high-impact protein domains tend to be extremely conserved, making it tricky to pinpoint specific positions. Variants also occur sparsely across the coding region and at very low frequencies within a population. Consequently, it is difficult to increase the number of variants for population analyses without increasing the pool of sequenced individuals. To this end, we devise an "intra-genome conservation" approach that is able to

"amplify" the variant signal in protein-coding regions within a population.

There is a wide range of repeat protein domains (RPDs) [10,11]. Each RPD is made up of modular repeat motifs of the same class. This modularity gives rise to a strategy for a particular class of RPDs that was first introduced in the field of protein engineering to generate protein design templates to create synthetic proteins with desired specificities and affinities [12–14]. We adapted the strategy to build a multiple sequence alignment (MSA) profile, which we term a "motif-MSA" profile, for each class of RPD. As an initial proof-of-concept for our novel approach, we focus on this category of RPDs that has been shown to be amenable to the motif-MSA approach. This category of RPDs explicitly mediates protein–protein interactions (PPIs), and their repeat motifs in each RPD require each other to maintain their structural fold. Each repeat unit is also relatively short with length of 12–100 aa. Many of these classes of RPDs have been studied extensively [15–17]. For example, tetratricopeptide repeat (TPR) domains are made up of only TPR motifs and Ankyrin repeat (ANK) domains of ANK repeat motifs. Using the TPR as an example of a class of PPI RPD, we demonstrate that the motif-MSA strategy can "amplify" variant signal by aggregating the variants from all homologous motifs for each class of RPD within the human genome. Interestingly, we note that such analyses of intra-genome conservation can only be performed using a dataset as large as those from the Exome Aggregation Consortium (ExAC) database [1]. Our Intensification database contains our results as a resource for annotating variants in 12 PPI RPDs (see "Methods" for selection criteria).

Results

Intensification database

Figure 1a shows our strategy that is used to build up the resources in our publicly available Intensification database[‡] that relates protein residue to genomic information in 12 RPDs, which encompass 5508 motifs and 971 proteins in *Homo sapiens* (Supplementary Table 1). Our strategy first produces a motif sequence alignment profile for a class of repeat domain. We obtain every repeat motif of a given amino acid length in the human proteome (typically the length with the most number of available motifs). We then perform an MSA of all the motifs (motif-MSA) to obtain a residue frequency table, which shows the percentage occurrence of each amino acid at each position in the motif. This table can then be translated into a sequence logo for better visualization. For each repeat motif, we then locate its genomic positions in the human genome. Subsequently, we map single nucleotide variants (SNVs) onto the genomic coordinates of the repeat motifs. This allows us to obtain aggregate counts of variants at each residue positions for each class of repeat domain based on SNV allele frequencies and the functional impact, namely whether the SNV is rare (R) or common (C) in the human population and whether the SNV causes a synonymous (S) or non-synonymous (NS) change. From these statistics, we can subsequently derive more meaningful metrics such as ratio of NS-to-S-SNV profile (NS/S) and enrichment of rare variants (R/C) for interpretation of each residue position. We provide these results for the users in our Intensification database. Here, we use the 34-aa TPR repeat motif as an example (see "Methods" for details; Fig. 1 and Supplementary Fig. 1).

Comparing species- and motif-MSA

An MSA is more typically performed using homologous sequences from multiple species (Fig. 1b; we term "species-MSA"). Here, we perform species-MSA for the first three TPR motif sequences in the TPR-containing protein TTC21B, using orthologous sequences from 66 species (see "Methods" for details;-Fig. 2a). TTC21B contains about 16–19 TPR motifs, with almost all of them having a length of 34 aa, and is a cilia-specific protein that is necessary for retrograde intraflagellar transport [18]. Expectantly, most positions are comparably high in sequence conservation. In contrast, the motif-MSA profile exhibits substantially

Fig. 1. Our motif-MSA approach amplifies variant information as compared to species-MSA. (a) (1) We first query a database and obtain all the proteins with the desired domains or motifs. We use the TPR motifs as an example in this figure. These motifs have to be of the same length. Here, we select TPR motifs that are 34 aa since it is the size that most frequently occurs. (2) Subsequently, we perform an "ungapped" multiple sequence alignment (MSA) of the human TPR motifs by lining them up end to end, in order to obtain a sequence conservation profile. This motif-based MSA (black sequence logo) typically exhibits differential sequence conservation among the positions across the length of the motif. (3) The third step involves collecting genomic single nucleotide variants (SNVs) for each amino acid position of the motif-based alignment profile. We make use of the corresponding genomic coordinates of the TPR motifs in the motif-MSA to aggregate variants over all motifs within the human genome, thereby amplifying variant information sufficiently for further downstream analyses. (4) For each motif-MSA, we then host the results on our Intensification database. For each protein repeat domain, we build a motif-MSA and compute corresponding SNV profiles, including residue frequency tables, log(NS/S), log(R/C), and SIFT score distributions. (b) For conventional species-MSA, orthologous sequences are typically aligned across multiple species. However, because we are only using the corresponding genomic coordinates of the protein in the human genome, only three variant positions can occur at each codon in a species-MSA profile.