Gene 550 (2014) 155–164

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Clustering of gene ontology terms in genomes

Timo Tiirikka [a,b,*], Markku Siermala [a,1], Mauno Vihinen [a,c]

[a] Institute of Biomedical Technology, University of Tampere, Finland and BioMediTech, FI-33014 Tampere, Finland
[b] Medical Research Center Oulu, Oulu University Hospital, University of Oulu, Finland
[c] Department of Experimental Medical Science, Lund University, SE-22 184 Lund, Sweden

ARTICLE INFO

Article history:
Received 9 May 2012
Received in revised form 26 June 2014
Accepted 27 June 2014
Available online 1 July 2014

Keywords:
Genomics
Bioinformatics
Systems biology
Computational biology

ABSTRACT

Although protein coding genes occupy only a small fraction of genomes in higher species, they are not randomly distributed within or between chromosomes. Clustering of genes with related function(s) and/or characteristics has been evident at several different levels. To study how common the clustering of functionally related genes is and what kind of functions the end products of these genes are involved, we collected gene ontology (GO) terms for complete genomes and developed a method to detect previously undefined gene clustering. Exhaustive analysis was performed for seven widely studied species ranging from human to *Escherichia coli*. To overcome problems related to varying gene lengths and densities, a novel method was developed and a fixed number of genes were analyzed irrespective of the genome span covered. Statistically very significant GO term clustering was apparent in all the investigated genomes. The analysis window, which ranged from 5 to 50 consecutive genes, revealed extensive GO term clusters for genes with widely varying functions. Here, the most interesting and significant results are discussed and the complete dataset for each analyzed species is available at the GOme database at http://bioinf.uta.fi/GOme. The results indicated that clusters of genes with related functions are very common, not only in bacteria, in which operons are frequent, but also in all the studied species irrespective of how complex they are. There are some differences between species but in all of them GO term clusters are common and of widely differing sizes. The presented method can be applied to analyze any genome or part of a genome for which descriptive features are available, and thus is not restricted to ontology terms. This method can also be applied to investigate gene and protein expression patterns. The results pave a way for further studies of mechanisms that shape genome structure and evolutionary forces related to them.

1. Introduction

Numerous complete genomes have been sequenced during the last decade. Because only a small fraction of each eukaryotic genome encodes proteins, genes have been thought to be randomly distributed within and between chromosomes. However, the organization of genes within eukaryotic genomes is clearly non-random (Hurst et al., 2004; Kosak and Groudine, 2004; Michalak, 2008). Notably, regions containing the most actively expressed genes have higher gene density (Versteeg et al., 2003; Woo et al., 2010). Consequently, regions of increased gene expression (ridges) are gene dense and have high G + C content (Versteeg et al., 2003). Serial analysis of gene expression (SAGE) and microarray studies have indicated that a large portion of co-expressed genes are clustered in specific areas of genomes (Elizondo et al., 2009; Singer et al., 2005), examples of which come from *Saccharomyces cerevisiae* (Cho et al., 1998; Cohen et al., 2000), *Drosophila melanogaster* (Boutanaev et al., 2002; Spellman and Rubin, 2002), *Homo sapiens* (Caron et al., 2001), *Gallus gallus* (Nie et al., 2010), *Caenorhabditis elegans* (Roy et al., 2002) and *Danio rerio* (Tsai et al., 2009). So called housekeeping or maintenance genes, which are expressed in most tissues, are also clustered (Lercher et al., 2002). In light of these results, genomes seem to be organized to facilitate efficient regulation of specific gene processes relating e.g. tissue formation (Al-Shahrour et al., 2010; Dewey et al., 2010). The clustering of mammalian imprinted genes is a prime example of non-random gene ordering in eukaryotes (Morison et al., 2005).

The analysis of expression data for yeast, fruit fly, worm, rat, mouse and human indicated that neighboring genes are likely co-expressed (Fukuoka et al., 2004). The proximity of a pair of genes has been used to predict gene functions (Raghupathy and Durand, 2009; Yanai et al., 2002). In bacteria, gene essentiality determines chromosome organization (Rocha and Danchin, 2003), and essential genes occur more frequently and are conserved in the leading replicating strand as compared with the expected average frequency for all genes. Protein sequences offer additional information about gene co-expression via

Abbreviations: GO, gene ontology; SAGE, serial analysis of gene expression; KEGG, Kyoto Encyclopedia of Genes and Genomes; MHC, major histocompatibility complex; OR, olfactory receptor; CC, cellular component; MF, molecular function; BP, biological process.

* Corresponding author at: Medical Research Center Oulu, Oulu University Hospital and, University of Oulu, Finland.
E-mail addresses: timo.tiirikka@student.oulu.fi (T. Tiirikka), markku.siermala@luukku.com (M. Siermala), mauno.vihinen@med.lu.se (M. Vihinen).
[1] Present address: Logica Oy, Hatanpäänkatu 3 F, PL 287, 33900 Tampere, Finland.

http://dx.doi.org/10.1016/j.gene.2014.06.060
0378-1119/

network maps using the "betweenness" concept as an indicator of proteins having interrelated functions (Yu et al., 2007).

Analysis of Kyoto Encyclopedia of Genes and Genomes (KEGG) metabolic and signaling pathways in five eukaryote model species revealed that a high proportion of genes for individual pathways are clustered in each species' genome (Lee and Sonnhammer, 2003). There are differences among the species; however, 30–98% of the genes in the 69 investigated pathways were clustered. Still, only seven of the pathways were clustered for all the eukaryotes studied.

Many functionally related genes are organized in bacteria in operons, and operon-like gene clusters have been identified in many species including e.g. plants, animals, and also human (Osbourn and Field, 2009). Gene duplications generate groups of related genes (for a review see Reams and Neidle, 2004). There are also other mechanisms, especially for clusters of non-homologous genes. As the extent of clustering has not been systematically investigated, we performed genome wide studies for several model organisms based on gene annotations.

Genes and genomes have been annotated in many ways. Gene ontology (GO) terms are rich annotations of function, components, and cellular localization (Ashburner et al., 2000). Previously, GO terms were examined in some of the co-expression clusters in human and yeast (Fukuoka et al., 2004). Certain clusters were identified, but the events were rather rare. In another study, the chromosomal locations of DNA binding proteins encoded on human chromosome 19 strongly correlated with GO annotations (Castresana et al., 2004). Stanley et al. (2006) developed a method to identify statistically significant GO terms associated with genomic positions. However, the number of genes and GO terms was relatively small in these studies.

The number of sequenced genomes is growing steadily. Although annotations have lagged behind, there are already a number of well annotated genomes with functional information for most genes and proteins. Here, we investigated the genome-wide GO distribution in seven species for which complete genomes are available, namely *H. sapiens*, *Mus musculus*, *S. cerevisiae*, *C. elegans*, *D. melanogaster*, *Arabidopsis thaliana* and *Escherichia coli*. We determined the GO terms for each gene and gene product and then calculated statistics for the enrichment of GO terms in adjacent genes. The results indicate clear clustering of GO terms and functions within chromosomes and sequence regions in all the investigated species, and certain features of GO term distributions appear to be species specific.

We developed a method to investigate the co-occurrence of ontology annotations in genomes. The method is based on statistical analysis and provides information for a fixed number of consecutive genes, which facilitates analysis independent of gene density. This feature is advantageous because e.g. in human less than 5% of the genome contains protein-coding genes, and gene density varies significantly for different chromosomes and regions in them. Previous genome-wide clustering studies have been restricted to a standard length of a studied genome region (Kano et al., 2003), which provides limited insight into the clustering phenomenon. In another approach, genome positions rather than genes were assumed to be randomly distributed (Stanley et al. 2006). The effect of different gene sizes was avoided; however, they did not perform a genome-wide GO term analysis. The C_Hunter program (Yi et al., 2007) is more similar to our approach; however, it focuses on finding the longest GO term clusters in studied species and does not further analyze its findings on the genome level. In addition, DEFOG, a web based application by Wittkop et al. (2012) uses a resembling way to organize genes in a pathway to functionally related units in order to reduce the complexity of the clustering task. However, our method, which avoids gene length and size bias, was used to analyze GO term distributions in numerous complete genomes.

2. Results and discussion

Our aim was to reveal how genes with related functions are distributed in genomes. The analysis is based on GO terms: a systematic description of molecular functions, biological processes, and cellular components. GO annotations were retrieved from NCBI Entrez Gene. GO terms are currently incomplete for any species, yet they are very useful and well suited for genome-wide statistical analyses. Some properties of the analyzed genomes are listed in Table 1. The number of genes varies widely, from 4279 to 38,699, among the species that we investigated, and the human genome contains the largest number of genes. Among these genomes, there are from 0.88 (*S. cerevisiae*) to 8.4 (mouse) GO terms per gene on average. The average human gene has 4.4 annotations, which is about half of that for mouse (8.4). The ratios of GO term classes are somewhat different for each species (Fig. 1). Cellular component is the smallest GO term category in all examined cases.

The analysis of the human genome was performed starting with 38,699 genes, which is higher than the number of current, officially named genes because the automated analysis is based on genome annotations. The GO term coverage, i.e., the percentage of genes for which GO terms were found, was 25.4%. Altogether there were 53,844 molecular function, 51,631 cellular component, and 65,760 biological process ontology terms, totaling 171,235 GO terms (Table 1).

In order to illustrate how syntenic regions, genes and other markers with an evolutionary conserved order localize with GO term distribution we analyzed human and mouse syntenic regions alongside with mouse GO term clusters. The results, chromosome wise, are in Supplementary Figs. 12 to 32. The vast majority of GO clusters and syntenic regions seem to follow each other verifying the biological clustering process. However, there are differences too.

2.1. Analysis method

Hypergeometric distribution was used for statistical tests because it works well even with small datasets. This test has been widely used for GO annotation distribution studies. The uncorrected p-value had to be 10^{-6} or lower for the results to be considered statistically significant. We were interested just on the most significant findings which were obtained with this p-value. In addition, Bonferroni correction for multiple testing and false discovery rate (FDR) were used to overcome statistical problems regarding multiple testing. As the outcome of the analysis was very similar for the two corrections, results are only shown for Bonferroni corrected data (Supplementary Tables 1 and 2). Every GO term and level was considered to be equally important so we did not use scoring methods (see e.g. Alexa et al., 2006) to weigh for more detailed terms. We slid a window of a fixed number of genes for each investigated chromosome. The window was moved in steps of one gene. The width of the window was set to five genes and increased in five-gene increments until 50 consecutive genes were included at one time. The reason for the choice of the range of window sizes was that based on published information widely different sizes of clustered genes had been identified. Our goal was to investigate the extent of the phenomenon of functionally related genes in diverse species. Statistics were calculated for the distribution of GO terms within the window. Because the number of genes with ontology term classes varies widely among species, we used expected values for the random occurrence of the GO terms. Results are calculated based on existing annotations and such are affected by any features affecting annotations. Even if the annotations are biased in some way it is likely that the annotations of related genes are affected quite similarly. In order to verify that the phenomenon of the clustering of gene ontologies and thus functions, processes and components is valid, a randomization study was conducted. However, initial hypothesis was not affected by the randomization results; the phenomenon was statistically much stronger than the simulated study. The results are first shown for the human genome, and trends and differences among other genomes are discussed in later chapters.

2.1.1. GOme database

To maintain and distribute vast amounts of data for different species, we created the GOme database. It is a mySQL-based system containing