Gene 515 (2013) 277–280

Contents lists available at SciVerse ScienceDirect

Gene

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

Exploring charged biased regions in the human proteome

Mouna Choura, Ahmed Rebaï *

Molecular and Cellular Diagnosis Processes, Centre of Biotechnology of Sfax, University of Sfax, Route Sidi Mansour, Po Box 1177, 3018 Sfax, Tunisia

ARTICLE INFO

Article history:
Accepted 28 November 2012
Available online 20 December 2012

Keywords:
Biased
Charged
Cluster
Human
Proteome
Residues

ABSTRACT

There has been an increasing interest in biased regions in proteins especially ever since it was shown that such regions are frequently associated with a structural role in the cell, or with protein disorder. In this study, we focus on charged biased protein sequences in human genome. We have identified 446 charged biased proteins within human proteome, 70% of them constitute proteins harboring negative run that correspond to transcription factor zinc finger proteins, importins and some protein kinases involving acidic activating domains. Basic charge clusters are often associated with DNA-binding, zinc-finger, basic-leucine zipper and homeobox domains. The data show that significant positive clusters correspond to ribosomal proteins. Most of proteins with zinc-binding fingers have a mixed positive and negative charged biased regions. Altogether, the Gene Ontology analysis revealed that the charged proteins are involved mainly in regulatory functions.

1. Introduction

Most protein sequences contain segments called compositionally biased (CB) regions, whose amino acid composition is significantly different from the average amino acid usage of the proteome. CB regions are sequence stretches with a large fraction of a small subset of residue types. If the CB regions are biased for multiple amino-acid residue types that are strewn along the sequence in an irregular way, the boundaries of these regions can be difficult to define. Statistical analyses indicate that up to 25% of the proteome can be composed of (CB) (Wootton and Federhen, 1996). It was shown that proteins involved in the same biological function tend to contain the same types of (CB) (Harrison, 2006; Kuznetsov and Hwang, 2006). Recent observations indicate that particular types of (CB) in proteins are often structurally disordered (Romero et al., 2004). These regions have also been linked to protein misfolding (DeMarco and Dagget, 2007) and a number of inherited neurological diseases in human (Gunawardena and Goldstein, 2005; Harrrison and Gerstein, 2003; Kreil and Ouzounis, 2003). Moreover, comparative proteome analysis of five complete eukaryotic genomes found that multiple amino acid runs are often associated with diseases; these include long glutamine runs that induce neurological disorders, various cancers, and categories of leukemia (Karlin et al., 2002).

Global compositional bias, when the entire protein sequence contains a large excess of particular amino acid types, is also known to be related to the general protein function. For instance, the net positive charge of histones that results from a bias in amino acid usage towards positively charged residues facilitates interactions with negatively charged DNA (Karlin et al., 2003). Previous study has identified and associated significant charged clusters defined as 25 to 75 residues with functional domains of cellular transcription factors (Brendel and Karlin, 1989).

Positive charge clusters are very rare. Negative or histidine-acidic charge clusters often coordinate calcium or magnesium or zinc ions. They are also involved in protein–protein interaction and in substrate binding (Zhu and Karlin, 1996).

Our objective here is to provide a more complete distribution and new insight into the role of charged biased protein sequences in the human proteome.

2. Methods

2.1. Dataset

Protein sequences were retrieved from LPS-annotate server http://cedra.biol.mcgill.ca/lps-annotate.html (Harbi et al., 2011) which consists of pre-calculated annotations of compositionally-biased (CB) in the SwissProt database (Release 2012). The database was searched for KR, RK, ED and DE biases. Only human proteins were selected for further analysis.

2.2. GO analysis

Gene functions were assigned according to Gene Ontology (GO) annotations (Ashburner et al., 2000), and the statistical significance of gene ontologies within a group of genes was analyzed by the GOstat program against all GO annotated human. p-Values were calculated using Fisher's exact test and then adjusted the p-values by

Abbreviations: CB, charged biased; DE, aspartic acid and glutamic acid; KR, lysine and arginine; GO, Gene Ontology.

* Corresponding author. Tel.: +216 74875818; fax: +216 74871816.
E-mail address: ahmed.rebai@cbs.rnrt.tn (A. Rebaï).

0378-1119/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.gene.2012.11.077

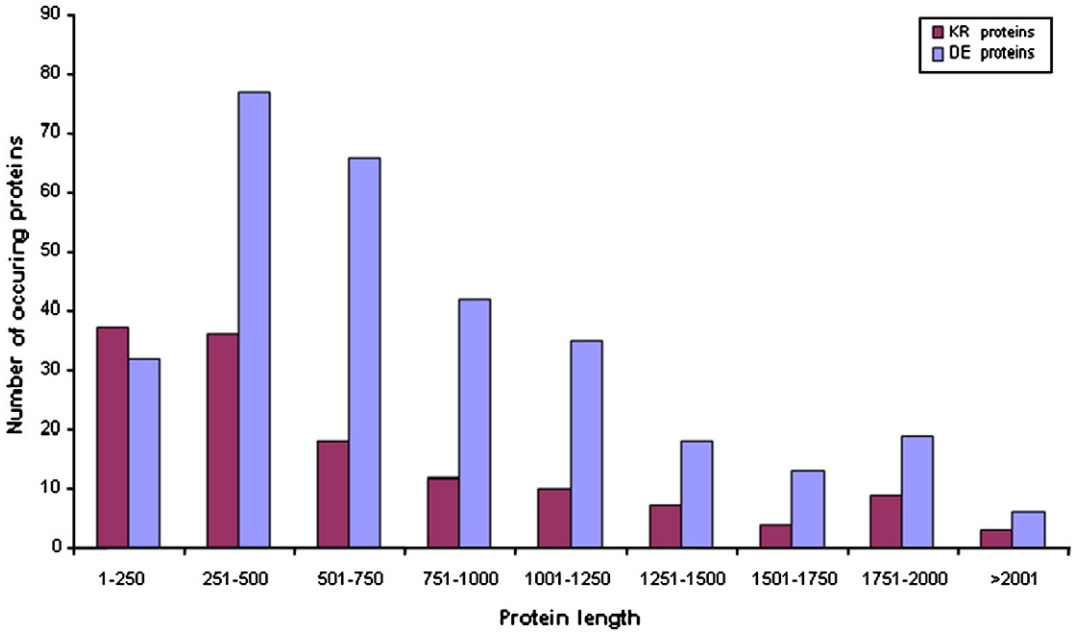

Fig. 1. Distribution of proteins containing charged biased regions (ED and KR clusters) depending on their length.

FDR using Benjamini–Hochberg procedure (http://gostat.wehi.edu.au/) (Beissbarth and Speed, 2004).

2.3. Functional classification

The gene functional annotation analysis of proteins was performed using DAVID bioinformatics resources at http://david.abcc.ncifcrf.gov/home.jsp (Sherman et al., 2007). We select the 'highest' stringency, giving group with high enrichment score. The enrichment score is the minus log scale of geometric mean of each member's Fisher exact p-values in that cluster (below the formula). A higher enrichment score indicates that gene members in the group are involved in more important (enriched) terms, therefore they deserve more attention.

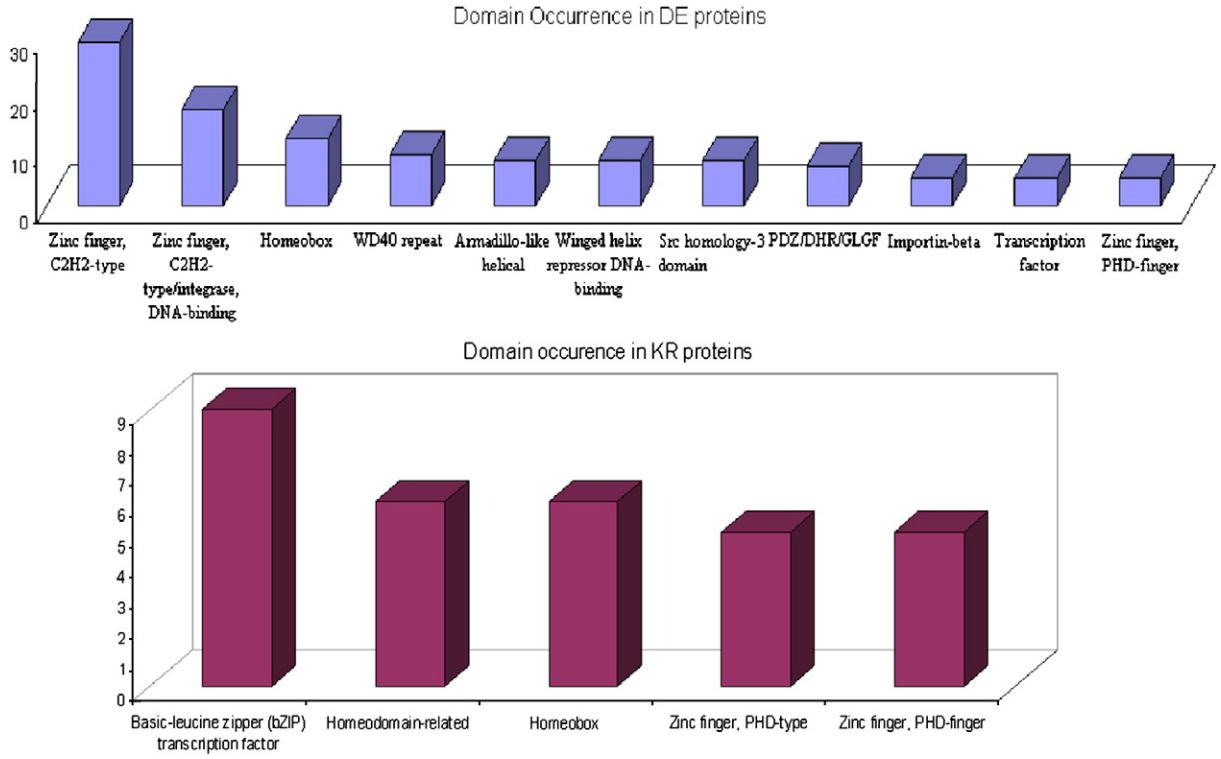

Fig. 2. Domain occurrences in DE proteins and KR proteins.