Available online at www.sciencedirect.com
SCIENCE DIRECT®
Mechanisms of Ageing and Development 126 (2005) 193–208
mechanisms of ageing and development
www.elsevier.com/locate/mechagedev

Text-based analysis of genes, proteins, aging, and cancer

Jeremy R. Semeiks, L.R. Grate, I.S. Mian*

Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

Available online 26 October 2004

Abstract

The diverse nature of cancer- and aging-related genes presents a challenge for large-scale studies based on molecular sequence and profiling data. An underexplored source of data for modeling and analysis is the textual descriptions and annotations present in curated gene-centered biomedical corpora. Here, 450 genes designated by surveys of the scientific literature as being associated with cancer and aging were analyzed using two complementary approaches. The first, ensemble attribute profile clustering, is a recently formulated, text-based, semi-automated data interpretation strategy that exploits ideas from statistical information retrieval to discover and characterize groups of genes with common structural and functional properties. Groups of genes with shared and unique Gene Ontology terms and protein domains were defined and examined. Human homologs of a group of known *Drosphila* aging-related genes are candidates for genes that may influence lifespan (*hep*/MAPK2K7, *bsk*/MAPK8, *puc*/LOC285193). These JNK pathway-associated proteins may specify a molecular hub that coordinates and integrates multiple intra- and extracellular processes via space- and time-dependent interactions with proteins in other pathways. The second approach, a qualitative examination of the chromosomal locations of 311 human cancer- and aging-related genes, provides anecdotal evidence for a ''phenotype position effect'': genes that are proximal in the linear genome often encode proteins involved in the same phenomenon. Comparative genomics was employed to enhance understanding of several genes, including open reading frames, identified as new candidates for genes with roles in aging or cancer. Overall, the results highlight fundamental molecular and mechanistic connections between progenitor/stem cell lineage determination, embryonic morphogenesis, cancer, and aging. Despite diversity in the nature of the molecular and cellular processes associated with these phenomena, they seem related to the architectural hub of tissue polarity and a need to generate and control this property in a timely manner.

Keywords: Statistical information retrieval; Progenitor/stem cell lineage determination; Embryonic morphogenesis; Cancer; Aging; Phenotype position effect

1. Introduction

Genomic and genetic approaches to the study of cancer and aging are widespread. An underutilized resource in investigations of these and other biological phenomena is the biomedical literature. Efforts to exploit data in the form of text seek to convert information into knowledge in a systematic and quantitative manner, most often using tools and techniques developed for modeling text documents in other scientific disciplines (reviewed in Shatkay and Feldman, 2003). For example, probabilistic graphical models have proved invaluable in the analysis of molecular sequence and profiling data, and recently this statistical framework has been applied to a corpus of documents about *C. elegans* in order to gain insights into aging (Blei et al., 2004).

A common outcome of biomedical research is the generation of a collection of ''interesting'' genes and/or proteins, for example, genes involved in response to stress, differentially expressed between normal and aged tissue samples, and so on. Here, the focus is two extant collections of genes that reviews of the scientific literature designated as being involved in cancer and in aging. This work illustrates how analysis of these collections using a combination of two general approaches currently underexploited in computational biology yields new and enhanced insights into these genes and phenomena. The first approach, ensemble attribute profile clustering, is the semi-automated functional grouping of genes and/or gene products based on their shared annotations from a set of terms specified in curated textual corpora. The second approach considers the order of

* Corresponding author.
E-mail address: smian@lbl.gov (I.S. Mian).

0047-6374/$ – see front matter © 2004 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.mad.2004.09.028

genes in a genome (as opposed to their actual sequences) and is based on the hypothesis that genes that are close in terms of linear order are involved in the same phenomenon (''phenotype position effect''). The integrated text-based analysis described here both recapitulates known properties of genes involved in the generalized phenotypes of cancer and aging, and suggests new candidates for genes associated with these phenomena.

Ensemble attribute profile clustering is a recently formulated strategy designed to assist in the analysis of a set of genes (Semeiks et al., 2004). Using data in the form of textual descriptions rather than molecular sequences or profiles, it addresses the task of discovering and characterizing groups of genes with common functional, structural, and other properties (''attributes''). The approach exploits ideas from statistical information retrieval, a field which, in contrast to Natural Language Processing, emphasizes vector space and probabilistic models for document representation, retrieval and analysis (Salton, 1988). Vector space models of documents ignore syntax and semantics when recasting a document as a bag of words. Similarly, gene attribute profiles, the vector space model representation of genes employed by ensemble attribute profile clustering, neglect the order in which domains occur in a protein, discard the manner in which terms in a biological ontology are organized, and so on when recasting a gene as a set of attributes. Next, simple probabilistic graphical models are estimated from a set of gene attribute profiles, the models used to define groups of genes with similar patterns of attributes, and the biology of genes assigned to groups examined by a user. Previously, this type of graphical model was employed to cluster transcript profiles and thus determine groups of genes (or experiments) with common patterns of expression (Moler et al., 2000a, b; Bhattacharjee et al., 2001). Here, analysis of 291 cancer- and 159 aging-related genes using attributes derived from the gene-centered LocusLink corpus and ensemble attribute profile clustering demonstrates the practical utility of this semi-automated data interpretation strategy.

Recent studies indicate that the distribution of genes along eukaryotic chromosomes is not random, that is, nearby genes may be co-expressed, be involved in the same metabolic pathway, interact with each other, and share regulatory regions, histone modification states, or regulatory elements (reviewed in Hurst et al., 2004). Most investigations pertaining to the biological consequences of gene order have examined evidence regarding the effect of a gene's genomic location on the molecular endpoint of transcription, that is, an expression position effect. This work postulates the existence of a phenotype position effect, the interplay between genome organization and cellular/tissue/organismal endpoints. Inspection of genes in human genomic regions containing known cancer- and aging-related genes suggests relationships between progenitor/stem cell lineage determination, embryonic morphogenesis, cancer, and aging. The roles of tissue polarity and cryptic genetic variation are discussed.

2. Methods

2.1. Known cancer- and aging-related genes

The two published collections of cancer- and aging-related genes reexamined here were derived from surveys of the scientific literature (cancer, Futreal et al., 2004), http://www.sanger.ac.uk/genetics/CGP/Census/; aging, March 2004 SAGEKE database (K. LaMarco, personal communication), http://sageke.sciencemag.org). The 291 cancer genes were exclusively human, whereas the 167 aging genes were from a range of species (*Homo sapiens*, 20 genes; *Mus musculus*, 18; *Drosophila melanogaster*, 30; *Caenorhabditis elegans*, 97; *Rattus norvegicus*, 2).

2.2. Ensemble attribute profile clustering

In order to discover and characterize groups of cancer- and aging-related genes with similar patterns of structural and functional properties, the two published collections were analyzed using ensemble attribute profile clustering. A comprehensive discussion of this general-purpose approach can be found elsewhere (Semeiks et al., 2004) so only a summary of the tasks involved is given below.

Data conversion maps gene identifiers in a source list to entries in a curated biomedical corpus. Here, the corpus employed was LocusLink (http://www.ncbi.nlm.nih.gov/LocusLink). For the July 2004 build of LocusLink, 291 of the cancer genes and 159 of the aging genes could be equated with distinct genetic loci, that is, LocusIDs. These two collections will be referred to as the Cancer and Aging data sets.

Feature generation represents each gene as a vector of (exchangeable) attributes where an attribute corresponds to a property of the gene and/or its protein product(s). Each element or ''feature'' of the vector signifies the significance or weight of the attribute in question in the gene of interest. Here, only protein-related attributes in general, and LocusLink-assigned Gene Ontology (GO; http://www.geneontology.org) (Ashburner et al., 2000) and Conserved Domain Database (CDD) (Marchler-Bauer et al., 2003) controlled vocabulary terms in particular, were considered. Examples of two attributes are the GO terms `transcription, DNA-dependent` and CDD term `Tyrosine kinase, catalytic domain. Phosphotransferases`.

Feature selection seeks to avoid overfitting the data during subsequent clustering by discarding uninformative attributes. Here, the problem of overfitting was reduced by retaining an attribute only if it was assigned to at least three of the N total genes in a set. If P is the number of attributes satisfying a selection criterion, the N P-dimensional vectors