Biochemistry and Biophysics Reports 1 (2015) 33–51

Contents lists available at ScienceDirect

Biochemistry and Biophysics Reports

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

# Intrinsically disordered proteins in the nucleus of human cells

Telma Frege [a,b], Vladimir N. Uversky [a,c,d,e,f,*]

[a] Department of Molecular Medicine, Morsani College of Medicine, University of South Florida, Tampa, FL 33612, USA
[b] GenomeNext LLC, 175 South 3rd Street, Suite 200, Columbus OH 43215, USA
[c] USF Health Byrd Alzheimer's Research Institute, Morsani College of Medicine, University of South Florida, Tampa, FL 33612, USA
[d] Department of Biology, Faculty of Science, King Abdulaziz University, P.O. Box 80203, Jeddah 21589, Saudi Arabia
[e] Institute for Biological Instrumentation, Russian Academy of Sciences, 142290 Pushchino, Moscow Region, Russia
[f] Laboratory of Structural Dynamics, Stability and Folding of Proteins, Institute of Cytology, Russian Academy of Sciences, St. Petersburg, Russia

ARTICLE INFO

*Article history:*
Received 8 February 2015
Accepted 11 March 2015
Available online 24 March 2015

*Keywords:*
Intrinsically disordered proteins
Cell nucleus
*Homo sapiens*
Protein function
Nuclear proteins
Nuclear compartments

ABSTRACT

Intrinsically disordered proteins are known to perform a variety of important functions such as macromolecular recognition, promiscuous binding, and signaling. They are crucial players in various cellular pathway and processes, where they often have key regulatory roles. Among vital cellular processes intimately linked to the intrinsically disordered proteins is transcription, an intricate biological performance predominantly developing inside the cell nucleus. With this work, we gathered information about proteins that exist in various compartments and sub-nuclear bodies of the nucleus of the human cells, with the goal of identifying which ones are highly disordered and which functions are ascribed to the disordered nuclear proteins.



## 1. Introduction

Being the first discovered cellular organelle, the nucleus, this membrane enclosed organelle found in eukaryotic cells, was described for the first time by the early microscopist Antonie van Leeuwenhoek (1632–1723). The nucleus is a key component of the eukaryotic cell since it is the "container" of its genetic information that serves as the "control center" of the cell, which is responsible for the storage of genetic information and coordination of gene expression [1–3]. The number of nuclei within a cell varies between the species from one and four, with one being the most common case. This organelle generally occupies about 6% of the total size of the cell. Among the most important functions ascribed to the nucleus are: storage of hereditary material (in chromosomes and genes); storage of proteins and RNA (specifically in the nucleolus); exchange of hereditary molecules (DNA and RNA); and production of ribosomes. The nucleus is a dynamic organelle, whose morphology (size and shape) is tightly regulated and is noticeably changed during the cell cycle [4]. There is a correlation between altered nuclear morphology and development of some diseases, e.g., cancer [4].

The cell nucleus is not a homogeneous entity, but contains several structures or compartments and sub-nuclear bodies [5]. Contrary to other components of the cell, most of these compartments are highly dynamic (do not exist all the time but only during certain stages of the cell, when those compartments are needed), and many of them are membrane-less, being formed via recruitment of proteins, RNA and DNA. Fig. 1 represents a schematic model of this membrane-enclosed organelle and shows that nucleus contains numerous nuclear domains or subnuclear organelles, such as nuclear pores, chromatin, nucleolus, PcG bodies (subnuclear organelles containing polycomb group proteins), Cajal bodies, promyelocytic leukemia (PML) nuclear bodies, Oct1/PTF/transcription (OPT) domains, nuclear speckles, nuclear gems (Gemini of coiled bodies), cleavage bodies, SAM68 nuclear bodies, perinuclear compartment, and several others [6]. Despite being different morphologically and functionally, all the aforementioned nuclear domains have some common features, e.g., all of them contain various types of RNA (or, in some cases, DNA) and different proteins.

Four levels of structural organization are known for a foldable ordered protein. Here, primary structure refers to the product of transcription, the protein amino acid sequence. Secondary structure corresponds to the 3D form of specific local segments, such as α-helix or β-strand. Tertiary structure represents the spatial conglomeration of secondary structure elements into a 3D superstructure. Tertiary structure is the highest form of the structural organization of a single-chain protein, whereas a multi-chain protein has quaternary structure that represents a more complex structural level constituting the assembly of tertiary structures.

Although for a long time it has been assumed that the presence of unique structure is a crucial prerequisite for protein to be

* Correspondence to: Department of Molecular, Medicine, Morsani College of Medicine, University of South Florida, 12901 Bruce B. Downs Boulevard, MDC07, Tampa, FL 33612, USA. Tel.: +1 813 974 5816; fax: +1 813 974 7357.
*E-mail address:* vuversky@health.usf.edu (V.N. Uversky).

http://dx.doi.org/10.1016/j.bbrep.2015.03.003

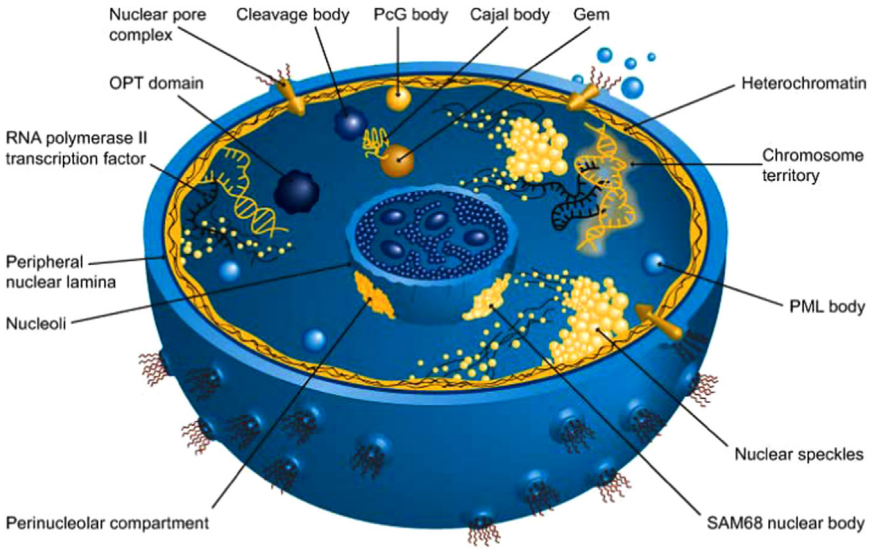


**Fig. 1.** Sub-nuclear compartments. Reproduced with the permission from [211].

functional, recent studied revealed that many biologically active proteins are characterized by a lack of tertiary structure. The discovery of these intrinsically disordered proteins (IDPs) and hybrid proteins consisting ordered domains and intrinsically disordered protein regions (IDPRs) challenged the protein structure paradigm stating that a protein must have a defined 3D-structure in order to perform a function [7–13]. Studies of different genomes suggested that IDPs are very abundant in nature and that proteins from eukaryotes have more intrinsic disorder than proteins of bacteria or archaea (up to 42% of all proteins in humans) [10,14–20].

Besides being very common in all analyzed proteomes, IDPs/IDPRs were shown to possess unique functional repertoire (being commonly involved in regulation, signaling and control pathways [21–23]), which is complementary to catalytic and transport functions of ordered proteins [24–27].

Since among the disorder-specific functions are DNA- and RNA-interactions, the goal of this work was to study the nuclear proteins of the human cell, to evaluate their level of disorder and to see if there is any sort of connection between intrinsic disorder and the functional roles these proteins play in the cell. The study is challenged by the fact that the information on components of the nucleus of human cells is scarce. Little information is currently available about the highly dynamic environment of this important cellular compartment, its organelles and interactions between them and the roles each such organelle plays. Furthermore, the molecular mechanisms defining the ability of these nuclear domains to maintain their specific structures in the absence of membranes also remain mostly unknown [28]. Studies of the proteins containing in each of these organelles (also called sub-nuclear domains or compartments) are sparse. Therefore, the overall goals of this project were: to find the proteins located in the human nucleus; to analyze distribution of these proteins within the sub-nuclear compartments; to analyze functions and structures of these proteins; to evaluate the intrinsic disorder propensities of these proteins; to look at the roles of intrinsic disorder in function and regulation of nuclear proteins; to map the distribution of disorder propensity within the nuclear compartments; and to study the relationship – if any – between the dynamic organelles within the cell nucleus and the level of disorder in the proteins that make them up.

## 2. Materials and methods

To solve this problem, a series of existing tools and data sources were used, and with them a pipeline was built to collect and process the information needed. Fig. S1 represents the resulting pipeline that had three major processing points, data collection, data processing, and data analysis.

### 2.1. Data collection

This stage had two steps. On the first step, the Nuclear Protein Database [29] was used to identify the proteins located inside the nucleus of eukaryotic cells. A total of 795 proteins were identified. On the second step, each protein was checked against UniProt to narrow the dataset to proteins that were curated and of human origin. The request for a protein to be curated provided us with the possibility to work only with proteins that have been reviewed and have reliable information. The protocol of protein curation in UniProt is shown in Fig. S1. The proteins of interest should be of human origin to filter out other species and focus on *Homo sapiens* only. The output of this stage was 185 proteins, collected in text files (FASTA sequences) and XML files. Each XML file contained all the details that UniProt provided for a given protein including all the names used for a protein, codes, cell location(s), functions, processes, sequence, etc. The distribution of these proteins in the different nucleus compartments is shown in Table 1. It is worth mentioning that the nucleus of the human cell has more proteins, but at the time this project started, the Nuclear Protein Database (http://npd.hgu.mrc.ac.uk/) [29] contained only 185 human curated proteins.

### 2.2. Data processing

On the second stage, which also had two steps, the 185 proteins were analyzed with a set of disorder predictors. There were several choices when it came to choose one, depending on the way the disorder is predicted. Because of their reputations and the detailed information they provide, the binary classifier CH-CDF plot and the PONDR-FIT® metapredictor [30] were chosen for this project.

#### 2.2.1. PONDR-FIT® processing

PONDR-FIT® is a protein disorder meta-predictor (combines several methods to predict the level of disorder of a given sequence). This tool has proven to be moderately more accurate than many other disorder predicting tools [30]. The input for PONDR-FIT® is the FASTA sequences of the proteins. The tool