Assessing Gene-Disease Relationship with Multifunctional Genes Using GO

Hisham Al-Mubaid
Dept. of Computer Science
University of Houston-CL
Houston, TX, USA
hisham@uhcl.edu

Mohamed Shenify
Dept. of Computer Science
Albaha University
Alaqiq, KSA
maalshenify@bu.edu.sa

Sultan Aljahdali
Dept. of Computer Science
Taif University
Taif, KSA
aljahdali@tu.edu.sa

***Abstract*—Multifunctional genes possess special significance due to their role in multiple biological and molecular activities in many organisms. In this paper, we study multifunctional genes in human with gene functional annotations in Gene Ontology, and we utilize the extensive information available for genes related to gene functionalities from the *Molecular Function* and *Biological Process* aspects of the Gene Ontology. Specifically, we examine the gene ontology annotations of human genes to assess the gene multifunctionality and their relationship with human diseases. We examined and analyzed all human genes having gene ontology annotations from GOA database and using the OMIM diseases. We combined all OMIM disease genes with the GO annotated human genes in relation to human diseases and the results of the relationships between highly annotated human genes and diseases are highly significant.**

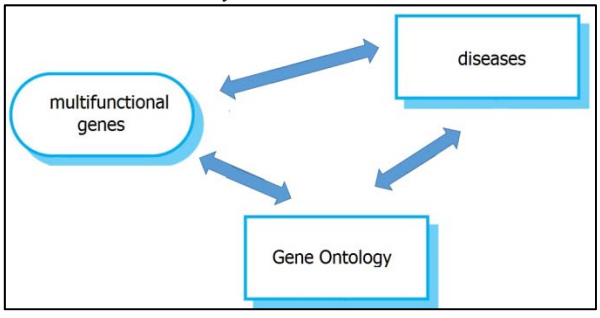
Keywords—multifunctional genes; gene ontology; gene-disease relationship.

1. INTRODUCTION

The study of human genes in relation to diseases counts as one of the very important biomedical subjects in the last few decades [1–3]. Many research projects have focused on investigating the gene functionalities, gene structure, and mutation in order to uncover most knowledge about genes. Gene functional analysis can reveal interesting discoveries about the roles of genes in the various molecular and biological activities in various organisms [2]. In human, it is important to discover and verify the gene disease associations and find the most knowledge about disease genes. Studies of gene activities and gene functional analysis in relation to diseases have led to significant findings and important results in the past two decades [1–4, 8–10]. These studies led to discovering multifunctional genes and proteins, also called moonlighting proteins. In this paper, we would like to investigate the *multifunctionality* of human genes from the Gene Ontology (GO) in relation with human diseases [5, 6].

In this work, we study the relationship between gene functionality and gene-disease relationship from a different prospective. In the medical and bioinformatics literature, there are quite a few studies and projects in the past ten years revealed a clear connection between diseases and genes having multiple functions [10–12]. Here, we investigate and analyze gene functionality using gene functional annotations (GOA) in GO. Gene Ontology is the official and most comprehensive database and taxonomy of gene functions and cellular component [5]. Gene ontology is composed of three aspects, or *sub-ontologies,* Molecular Function *mf*, Biological Process *bp*, and Cellular Component *cc* [5-6]. We use annotation terms in the *mf* and *bp* aspects of the GO, as a measure or indicator of multifunctionality of genes (Figure 1). In other words, we utilize the extensive information available about human genes in the GO annotation (*GOA*) databases to study the multifunctionality of human genes. Specifically, we examine the gene ontology annotations of genes to assess their multifunctionality and their relationship with human diseases. We examined and analyzed all human genes having gene ontology annotations (Table 1) from GOA database and using the OMIM diseases [7] as shown in Figure 1. The contribution here is to examine the relationship between gene multifunctionality and human diseases using the GO annotations as our multifunctionality indicator of human genes. We combined all OMIM disease genes (Table 2) with the GO annotated human genes in relation to human diseases (Figure 1) and the results are definitely significant. In other words, the examined human genes involved in many molecular and cellular processes as extracted from the GOA database were compared and analyzed with disease genes and found the relationship is significant.

Fig.1. The relationship between the Gene Ontology, gene multifunctionality and diseases.



978-1-5090-4320-0/16/$31.00 ©2016 IEEE

2. Background and Related Work

A number of research studies have reported in the past decade that multifunctional genes tend to have more association with diseases compared with non-multifunction genes [1–4, 8–10]. In other words, there is a signification relationship between diseases and multifunctional genes [1, 10]. When a protein-producing gene is involved in many activities, including molecular and cellular tasks, it can be called multifunctional gene, multifunctional protein, multitasking, or moonlighting protein [1-3, 8]. The main goal of studying multifunctional genes and proteins is basically to reveal more knowledge about diseases associated with the multifunctional genes and proteins. In this work we rely on the gene ontology which is the most popular repository of functional information about human genes [5, 6]. In [1], Pritykin, Ghersi, and Singh (2015) conducted a comprehensive study of genome-wide multifunctional genes in human. They found that multifunctional genes are significantly more likely to be involved in human disorders [1]. Also, they found that 32% of all multifunctional genes produced by their method are involved in at least one OMIM disorder, whereas the fraction of other annotated genes involved in at least one OMIM disorder is 21%.

This study is fairly comprehensive [1], and in fact, they found that, as compared to other annotated genes, genes involved in multiple biological processes or multiple molecular functions possess distinct physicochemical properties, are more broadly expressed, tend to be more central in protein interaction networks, tend to be more evolutionarily conserved, and are more likely to be essential [1].

Salathe et al. [12] studied the multifunctionality of genes and proteins in yeast for a different goal. They found a positive correlation between how many biological process (*bp*) GO terms a gene is annotated with and its evolutionary conservation in yeast; that is, they found highly significant negative correlation between number of *bp* GO terms and rate of change of yeast genes [12]. Also in [1], they observed that that multifunctional genes tend to be more evolutionarily conserved.

Another study of multifunctional genes [13] found some statistically significant positive correlation between the number of GO biological process leaf terms a gene has and its number of *Pfam* domains and are usually longer [13].

In [3], Khan et al. (2013) proposed a method for identification of novel moonlighting proteins from current functional annotations in public databases, even when they are not explicitly annotated as such. They identified potential moonlighting proteins in the Escherichia coli K-12 genome by examining clusters of GO term annotations taken from UniProt and constructed three datasets of experimentally confirmed moonlighting proteins [3].

3. Gene Functional Analysis

We investigated the gene functions for all human genes using gene ontology annotations database GOA from the Gene Ontology consortium [5, 6]. We downloaded the GOA gene-association for human which includes 490,198 annotations for 46,326 human genes (*Jan.2016*) in the three aspects *mf*, *bp*, and *cc* [6]. We also used the human disease database from the Online Mendelian Inheritance in Man OMIM data which includes diseases and genes related to human *OMIM* diseases; see Table 2. We found that 77.5% (*or 35893 genes*) of all human genes (= *46,326*) in the GOA database have *molecular function* annotations in the *human goa* association data; see Table 1. We investigated the relationship between disease genes and GO annotations of all human genes in the *mf* and *bp* aspects of the GO. It has been shown that disease genes in the human genome tend to have multiple functions [1–3, 10]. We would like to verify the multifunctionality characteristic of human genes using the GO in relation with human genes from OMIM disease database.

A gene annotated with multiple function terms in the GO (i.e., multiple GO *mf* terms) can be a good candidates of multifunctional genes thus can be a good candidate disease

TABLE 1: NUMBER OF HUMAN GENES AND GO ANNOTATION TERMS IN THE GOA DATABASE.

	MF	**BP**	**MF and BP**	**MF, BP, and CC**
Total annotations	154267	187029	341296	490198
Average annotations per gene	4.3	5.2	8.4	10.6
No. of records (genes)	35893	35732	40500	46326
No. of genes with 1 term	15142 (42.2%)	14660 (41.0%)	6164	7953
No. of genes with 2 terms	8278 (23.1%)	6516	9673	6837
No. of genes with 4 or fewer terms	28904 (80.5%)	25785 (72.2%)	24928 (61.6%)	26593 (57.4%)
No. of genes with 10 or fewer terms	32985 (91.9%)	31314	32456	35522 (76.7%)
No. of genes with 20 or more terms	1255 (3.5%)	1858 (5.2%)	4114 (10.2%)	6442 (13.9%)

TABLE 2. PHENOTYPES AND GENES IN OMIM DATABASE.

Human diseases and genes	
Total number of records	7062
Unique genes in *morbid map*	4837
Unique diseases	6640
Avg no. of genes per disease	1.37

TABLE 3. HUMAN GENES WITH THE HIGHEST NUMBER OF MF AND PB (COMBINED) ANNOTATIONS WITH THEIR ASSOCIATED DISEASES.

Human genes (*UniProtKB Id*)	***Number of GO annotation terms***			**Associated disease** (*Omim Id and disease name*)
	MF	**BP**	**MF + BP**	
P04637	420	142	562	*133239 ESOPHAGEAL CANCER*
P00533	345	73	418	*211980 ICD+ LUNG CANCER*
P62993	337	37	374	--
P35222	213	151	364	*114500 COLORECTAL CANCER; CRC*
Q08379	291	25	316	--
Q5S007	213	99	312	*607060 PARKINSON DISEASE 8, AUTOSOMAL DOMINANT; PARK8*
Q6A162	306		306	--
Q96EB6	136	137	273	--
Q12933	214	50	264	--
O60260	109	148	257	*168600 PARKINSON DISEASE, LATE-ONSET; PD*
P12931	139	117	256	--
P01137	46	199	245	*131300 CAMURATI-ENGELMANN DISEASE; CAEND*
Q00987	175	68	243	*614401 ACCELERATED TUMOR FORMATION, SUSCEPTIBILITY TO; ACTFS*
Q04206	153	87	240	*137800 GLIOMA SUSCEPTIBILITY 1; GLM1*
P0CG48	102	136	238	--
P10415	102	134	236	*151430 B-CELL CLL/LYMPHOMA 2; BCL2*
P31749	92	143	235	*114500 COLORECTAL CANCER; CRC*
Q7Z3S9	229	3	232	--
P27986	159	61	220	*615214 AGAMMAGLOBULINEMIA 7, AUTOSOMAL RECESSIVE; AGM7*
Q9UMX0	207	13	220	--

gene. We found that on average, a human gene is annotated with 4.3 *mf* GO terms, 5.2 *bp* GO terms and 8.4*mf* and *bp* GO terms combined as shown in Table 1. Therefore, a gene annotated with more than one *mf* or more than one *bp* terms may not necessarily be multifunctional gene. However, we found that more than 40% of human genes are annotated with only one *mf* term thus can be considered *non-multifunctional* as shown in Table 1 (same is true for *bp*). Moreover, our study showed that *OMIM* human diseases (from *morbid map*) have on average 9.3 *mf* GO term annotations.

We analyzed human GOA database [5, 6] and found that 90.9% (or 32,642) of human genes have 9 or fewer *mf* annotations (see Table 1) and this results is significant ($p<10^{-30}$) given that *morbid map* includes 6640 *OMIM* diseases. We also analyzed genes with 20 or more *mf* annotations (1255 genes or 3.5% of all *mf* annotated genes) and found ~ 66% of these genes are associated with disease (*OMIM morbid map*). This indicates that the relationship
between *multi-GO-mf* genes and diseases is significant ($p<10^{-30}$).
Next, for genes with >20 *bp* terms (1858 genes or 5.2% of all *bp* annotated genes) we got results of around 57% of these genes are associated with OMIM diseases and this is significant ($p<10^{-30}$). For genes having 20 or more annotation from *mf* and *bp* combined, we obtained around 51% of them associated with diseases (($p<10^{-30}$). These

TABLE 4: HUMAN GENES WITH THE HIGHEST NUMBER OF *MF*, *BP*, AND *CC* (COMBINED) ANNOTATIONS.

Human Genes (*UniProtKB Id*)	**Number of GO annotation terms**			
	MF	**BP**	**CC**	**MF + BP + CC**
P04637	420	142	67	629
P62993	337	37	226	600
P00533	345	73	126	544
P0CG48	102	136	284	522
P35222	213	151	115	479
P62979	7	150	311	468
P62987	4	152	300	456
P0CG47	14	151	287	452
P12931	139	117	126	382
Q5S007	213	99	49	361
Q08379	291	25	23	339
P27986	159	61	116	336
P31749	92	143	78	313
Q12933	214	50	49	313
Q6A162	306		2	308
P01137	46	199	59	304
Q96EB6	136	137	29	302
O60260	109	148	37	294
Q9Y4K3	101	85	101	287
Q04206	153	87	46	286

results yield that *mf* and *pb* annotations give significant indication about disease genes. The more a gene is annotated with *mf* or *bp* terms the more it likely is associated with diseases. Table 3 includes the results of the top 20 human genes with the highest GO annotations from both *mf* and *bp* combined and found that 11 of these 20 genes (or 55%) are associated diseases according to OMIM database; and this results is statistically significant ($p < 10^{-7}$). Table 4 contains number of GO annotations for the top 20 human genes with the highest number of *mf*+*bp*+*cc* annotations. Table 5 includes a small part of OMIM disease data showing gene symbols and number of GO terms annotating each gene; note each record (each row in Table 5) includes only one gene while each some diseases may be associated with more than one gene.

4. Discussion and Conclusions

When certain human gene produces a protein that is involved in some molecular functions or biological processes then such a gene is crucial for those molecular or cellular processes. Therefore, any *malfunctionality* in that protein-producing gene will lead to some disruption or commotion in that corresponding molecular or cellular process and may eventually lead to diseases. Consequently, genes involved in many molecular or cellular processes will be more susceptible to causing diseases than genes involved in one or very few molecular or cellular processes. In this work, we investigated human genes involved in many molecular and cellular activities using the gene ontology compared with disease genes and found the relationship is significant.

As it is shown in Table 3, among the 20 human genes with the highest *mf* and *bp* annotations, 11 of them are associated with OMIM diseases and therefore the relationship between multiple *mf* and *bp* annotations and diseases is significant with $p < 2 * 10^{-7}$. The top 20 human genes having the most annotations from GO by combining the 3 aspects (*mf*+*bp*+*cc*) are shown in Table 4.

Human gene BRCA2 (UniProtKB: P51587) is known to be the main gene for breast cancer disease. This gene (BRCA2_HUMAN, Breast cancer type 2 susceptibility protein BRCA2, OMIM: 600185, alternative symbol: FANCD1) is annotated with 8 distinct GO *mf* terms as shown in Table 6, indicating that this gene is involved in 8 different molecular functions; in fact it has a total of 48 GO *mf* annotations with 8 distinct mf terms.

We analyzed the relationship among these 8 terms in the GO *mf* aspect and the result is shown in Figure 2. It is clear that the two functions *protease binding* (GO:0002020) and *H4 histone acetyltransferase activity* (GO:0010485) are fairly distant (*shortest path length = 14*) in the ontology hierarchy (shown in Figure 2) suggesting that this gene (BRCA2) which is annotated with these two distant terms (GO:0002020 and GO:0010485) is highly likely multifunctional gene and thus associated with diseases.

We further found that this gene is annotated with 44 distinct *bp* terms (with a total of 53 *bp* annotations) which is another strong indicator that this gene is multifunctional and therefore it will be highly likely associated with one or more diseases. Table 7 contains the 8 diseases associated with this gene (*we did not include the bp terms and tree view of bp terms of BRCA2 for space limitation*). From this gene with Table 6 and Figure 2 we can draw a simple inference that if a gene is annotated with fairly many distinct GO *mf* (or *bp*) terms, then it is highly likely that we can find two of these terms fairly distant (far apart) from each other in the ontology hierarchy indicating that these two terms are semantically distant and representing strictly different functionalities; hence that gene is multifunctional.

Conclusion: We studied the multifunctionality of human genes in the Gene Ontology using the GOA *mf* and *bp* annotation terms. The GO *molecular function* terms annotated for a gene are basically the functions that the gene is involved in; and so multiple *mf* terms indicate that this gene might be involved in multiple molecular functions hence can be multifunctional gene. We showed that human genes highly annotated with *mf* or *bp* terms tend to be more likely associated with diseases according to OMIM disease database. In the future plans of this research, we would like to investigate the path length, depth, and information content of GO terms in the *mf* and *bp* aspects as an indicator of the multifunctionality of genes and its relationships to diseases.

Fig.2. Tree view of the eight *mf* terms annotating BRCA2 gene in the Molecular Function aspect.

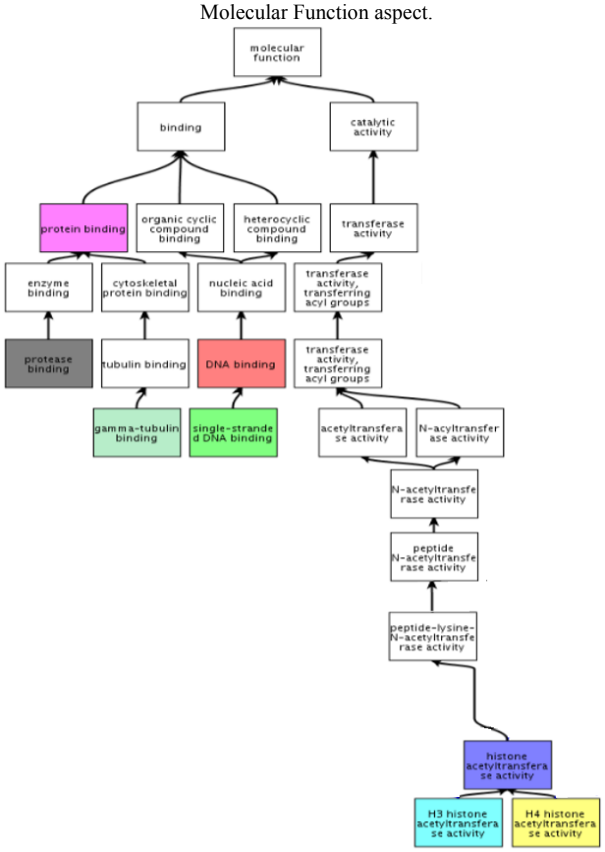


TABLE 5. SAMPLE OF OMIM DISEASE DATA INCLUDING DISEASE NAMES, ASSOCIATED GENE, AND NUMBER OF GO *MF* ANNOTATIONS FOR EACH GENE.

Phenotype (*disease name, disease OMIM Id*)	**Gene Symbols**	**OMIM Id**	**No. of MF terms**
Alcohol sensitivity, acute, 610251	ALDH2	100650	5
ACAT2 deficiency, 614055	ACAT2	100678	3
Multiple pterygium syndrome, lethal type, 253290	CHRNA1, ACHRD, CMS1B, CMS1A	100690	6
Myasthenic syndrome, congenital, 2C, associated with acetylcholine receptor deficiency, 616314	CHRNB1, ACHRB, SCCMS, CMS2A, CMS2C	100710	8
Myasthenic syndrome, congenital, 3A, slow-channel, 616321	CHRND, ACHRD, SCCMS, CMS3A, CMS3B, CMS3C	100720	4
Myasthenic syndrome, congenital, 4A, slow-channel, 605809	CHRNE, SCCMS, CMS4A, CMS4B, CMS4C	100725	5
Acne inversa, familial, 3, 613737	PSEN1, AD3	104311	29
Renal tubular dysgenesis, 267430	ACE, DCP1, ACE1, MVCD3, ICH	106180	29
Immunodeficiency 27A, mycobacteriosis, AR, 209950	IFNGR1, IMD27A, IMD27B	107470	5
17-beta-hydroxysteroid dehydrogenase X deficiency, 300438	HSD17B10, HADH2, ERAB, MRXS10	300256	15
Metacarpal 4-5 fusion, 309630	FGF16, MF4	300827	2
2-methylbutyrylglycinuria, 610006	ACADSB, SBCAD	600301	5
Mesothelioma, somatic, 156240	WT1, NPHS4	607102	23
3-Methylcrotonyl-CoA carboxylase 1 deficiency, 210200	MCCC1, MCCA	609010	8
17,20-lyase deficiency, isolated, 202110	CYP17A1, CYP17, P450C17	609300	7
Cerebrooculofacioskeletal syndrome 1, 214150	ERCC6, CKN2, COFS1, CSB, ARMD5, UVSS1	609413	21
3-M syndrome 1, 273750	CUL7, 3M1	609577	12
Deafness, autosomal recessive 28, 609823	TRIOBP, KIAA1662	609761	4
Albinism, oculocutaneous, type VI, 113750	SLC24A5, NCKX5, SHEP4, OCA6	609802	6
Amelogenesis imperfecta, type IIA5, 615887	SLC24A4, NCKX4, SHEP6, AI2A5	609840	6
COACH syndrome, 216360	TMEM67, MKS3, JBTS6, NPHP11	609884	6
3-M syndrome 2, 612921	OBSL1, KIAA0657, 3M2	610991	4
2-aminoadipic 2-oxoadipic aciduria, 204750	DHTKD1, KIAA1630, AMOXAD, CMT2Q	614984	2

TABLE 6. *MF* ANNOTATION TERMS FROM THE GO FOR BRCA2 GENE (UNIPROTKB: P51587).

Molecular Function (UniProtKB: P51587)	GO:0003677	DNA binding
	GO:0005515	protein binding
	GO:0003697	single-stranded DNA binding
	GO:0004402	histone acetyltransferase activity
	GO:0043015	gamma-tubulin binding
	GO:0002020	protease binding
	GO:0010485	H4 histone acetyltransferase activity
	GO:0010484	H3 histone acetyltransferase activity

TABLE 7. *BRCA2* GENE IS ASSOCIATED WITH 8 DISEASES IN OMIM.

Gene	Phenotype	Phenotype MIM number
UniProtKB: P51587 Breast cancer type 2 susceptibility protein BRCA2. OMIM: 600185	Fanconi anemia, complementation group D1	605724
	Wilms tumor	194070
	{Breast cancer, male, susceptibility to}	114480
	{Breast-ovarian cancer, familial, 2}	612555
	{Glioblastoma 3}	613029
	{Medulloblastoma}	155255
	{Pancreatic cancer 2}	613347
	{Prostate cancer}	176807

REFERENCES

[1] Y. Pritykin, D. Ghersi, M. Singh. Genome-Wide Detection and Analysis of Multifunctional Genes. PLOS Computational Biology, October 5, 2015

[2]. van de Peppel J, Holstege FCP (2005) Multifunctional genes. Molecular Systems Biology 1: 1–2. doi: 10.1038/msb4100006, 2005.

[3] I. Khan, Y. Chen, T. Dong, Hong X, Takeuchi R, et al. (2014) Genome-scale identification and characterization of moonlighting proteins. Biology Direct 9: 30. doi: 10.1186/s13062-014-0030-9 PMID: 25497125, 2014.

[4] Becker E, Robisson B, Chapple C, Guénoche A, Brun C (2012) Multifunctional proteins revealed by overlapping clustering in protein interaction network. Bioinformatics 28: 84–90. doi: 10.1093/ bioinformatics/btr621 PMID: 22080466, 2012.

[5] Ashburner et al. Gene Ontology: tool for the unification of biology (2000) *Nat Genet* 25(1):25-9. Online at Nature Genetics.

[6] The Gene Ontology Consortium. Gene Ontology Consortium: going forward. (2015) *Nucl Acids Res* 43 Database issue D1049–D1056. Online at Nucleic Acids Research, 2015.

[7] Online Mendelian Inheritance in Man, OMIM. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD), May 2016. World Wide Web URL: http://omim.org/

[8] Hernández S, Ferragut G, Amela I, Perez-Pons J, Pinol J, et al. (2014) MultitaskProtDB: a database of multitasking proteins. Nucleic Acids Research 42: D517–D520. doi: 10.1093/nar/gkt1153. pmid:24253302, 2014.

[9] Mani M, Chen C, Amblee V, Liu H, Mathur T, et al. (2015) Moonprot: a database for proteins that are known to moonlight. Nucleic Acids Research 43: D277–D282, pmid:25324305, 2015.

[10] A. Day, J. Dong, V. A. Funari, B. Harry, S.P. Strom, D.H. Cohn, and S. F. Nelson. Disease Gene Characterization through Large-Scale Co-Expression Analysis. PLoS ONE Vol.4, Issue 12, 2009.

[11] J. Gillis, P. Pavlidis (2013) Assessing identity, redundancy and confounds in gene ontology annotations over time. Bioinformatics 29: 476–482. doi: 10.1093/bioinformatics/bts727, 2013.

[12] M. Salathe M, Ackermann M, Bonhoeffer S (2006) The effect of multifunctionality on the rate of evolution in yeast. Molecular Biology and Evolution 23: 721–722, 2006.

[13] W.T. Clark, Radivojac P (2011) Analysis of protein function and its prediction from amino acid sequence. Proteins: Structure, Function, and Bioinformatics 79: 2086–2096, 2011.

[14] S. Hernandez et al. (2011) Do moonlighting proteins belong to the intrinsically disordered protein class? Proteomics Bioinformatics., 5, 262–264, 2011.

[15] S. Hernandez et al. (2014) MultitaskProtDB: a database of multitasking proteins. Nucleic Acids Res., 42, D517–D520, 2014.