

# Identification of Moonlighting Proteins in Genomes Using Text Mining Techniques

Aashish Jain, Hareesh Gali, and Daisuke Kihara\*

Moonlighting proteins is an emerging concept for considering protein functions, which indicate proteins with two or more independent and distinct functions. An increasing number of moonlighting proteins have been reported in the past years; however, a systematic study of the topic has been hindered because the secondary functions of proteins are usually found serendipitously by experiments. Toward systematic identification and study of moonlighting proteins, computational methods for identifying moonlighting proteins from several different information sources, database entries, literature, and large-scale omics data have been developed. In this study, an overview for finding moonlighting proteins is discussed. Then, the literature-mining method, DextMP, is applied to find new moonlighting proteins in three genomes, *Arabidopsis thaliana*, *Caenorhabditis elegans*, and *Drosophila melanogaster*. Potential moonlighting proteins identified by DextMP are further examined by a two-step manual literature checking procedure, which finally yielded 13 new moonlighting proteins. Identified moonlighting proteins are categorized into two classes based on the clarity of the distinctness of two functions of the proteins. A few cases of the identified moonlighting proteins are described in detail. Further direction for improving the DextMP algorithm is also discussed.

In function annotation, it is generally assumed that a protein has a single function and the possibility of the protein having an additional function is not extensively examined. However, over the past decade, an increasing number of proteins are accumulated that perform two independent and distinct functions. A classic example is an enzyme, L-argininosuccinate arginine-lyase, which was found to function as a lens structural protein delta-crystallin as well.<sup>[2]</sup> Proteins that have two functions have been called in several different ways, such as bi-, dual-, multifunctional proteins, multitasking proteins, gene sharing, promiscuous enzymes,<sup>[3]</sup> and moonlighting proteins,<sup>[4]</sup> but the latter two have rather specific definitions. A promiscuous enzyme is a protein that catalyzes a side reaction in addition to its main reaction. Moonlighting proteins perform two or more independent and distinct functions. In its original strict definition by Constance Jeffery, who coined the term,<sup>[4]</sup> the multiple functionalities are not due to gene


## 1. Introduction

Most molecular level studies in modern biology concern the functions of proteins and the mechanisms of how proteins carry out those functions. Thus, function annotation of proteins serves as fundamental information for biological studies. Algorithms for protein function prediction are extensively studied in bioinformatics.<sup>[1]</sup>

fusions, multiple domains, multiple splice variants, proteolytic fragments, families of homologous proteins, or pleiotropic effects. Some mechanisms identified to be responsible for the switch between two functions include different cellular localization of the protein, expression in different cell types, ligand binding sites, oligomerization states, and ligand concentration.<sup>[4]</sup> Many known moonlighting proteins were originally identified as enzymes, which were later found to have an additional function, such as transcription factors.

Moonlighting proteins that exhibit multiple functions can provide a competitive advantage to an organism from an evolutionary standpoint, especially in prokaryotes, where growth and the reproductive rate is directly associated with the number of genes translated and replicated.<sup>[4]</sup> Moonlighting proteins are also known to manage cellular activities by providing a coordinated framework by either self-regulation, e.g., thymidylate synthase, an enzyme that can bind to its own mRNA inhibiting its translation,<sup>[5]</sup> or by regulating other similarly functioning proteins, e.g., cystic fibrosis transmembrane conductance regulator, a chloride channel that also regulates epithelial sodium channel.<sup>[6]</sup> It was found that several moonlighting proteins play important roles in cellular activities that are involved in cancer and other diseases.<sup>[7]</sup> Thus, moonlighting proteins may be

Prof. D. Kihara  
Department of Biological Science  
Purdue University  
West Lafayette, IN, 47907, USA  
E-mail: dkihara@purdue.edu  
A. Jain, H. Gali, Prof. D. Kihara  
Department of Computer Science  
Purdue University  
West Lafayette, IN, 47907, USA  
Prof. D. Kihara  
Department of Pediatrics  
University of Cincinnati  
Cincinnati, OH, 45229, USA

 The ORCID identification number(s) for the author(s) of this article can be found under <https://doi.org/10.1002/pmic.201800083>

DOI: 10.1002/pmic.201800083

interesting drug targets to effectively suppress disease development if both functions of the proteins are involved in the target disease. On the other hand, blocking the activity of a moonlighting protein needs extra caution so that drugs only affect the desired function of the protein. Understanding the functional mechanisms of moonlighting proteins may lead to novel ideas for artificial design of proteins of dual function. It will also provide a foundation on how to avoid unexpected toxicity of artificially designed proteins and a protein artificially placed in a different cellular environment. With moonlighting proteins in the picture, our understanding of the functional interplay of proteins in a cell would need a major and fundamental update.<sup>[8]</sup>

Most of the currently known moonlighting proteins have been found serendipitously, where researchers identify a known protein as having a different function in an unrelated biological context. Jeffery's Lab manually compiled a list of known moonlighting proteins from literature in a database named MoonProt.<sup>[9]</sup> Multiple functions for the proteins in this database were reviewed by the authors based on published biochemical, mutagenic, and other evidence. There is another database, MultitaskProtDB-II,<sup>[10]</sup> where the authors curated a list of proteins that were found in PubMed with keywords indicating multiple functions: moonlight proteins, moonlighting proteins, multitask protein, multitasking proteins, moonlight enzymes, moonlighting enzymes, and gene sharing. Considering that we still only know a small number of moonlighting proteins, it is important to develop computational approaches that can systematically identify moonlighting proteins.<sup>[11]</sup> It was examined whether moonlighting proteins exhibit sequence similarity to protein families of different functions.<sup>[12]</sup> A second approach is to determine if there is a correlation between disordered regions and multifunctionality of proteins as disordered regions are often responsible for binding different proteins.<sup>[13]</sup> Another approach is to use protein–protein interaction (PPI) based on the idea that moonlighting proteins tend to interact with proteins with different functions or pathways reflecting their dual functionality.<sup>[14]</sup> Recently, Cheng et al. developed MoonFinder, which finds moonlighting long noncoding RNAs using subcellular location and function annotation of interacting proteins with long noncoding RNAs.<sup>[15]</sup>

Previously, our group has developed a framework of three methods for identifying potential moonlighting proteins based on the different types of information available about the proteins (Figure 1). Identifying moonlighting proteins on a large scale is a challenge even for cases when the two or more functions and their mechanisms are well known for proteins because the UniProtKB database<sup>[16]</sup> does not label such proteins with a specific keyword, e.g., moonlighting proteins or dual functional proteins. Thus, the right branch of the diagram in Figure 1 deals with cases where the dual functionality of proteins is known. When a protein's function is well studied, documented, and annotated with the gene ontology (GO)<sup>[17]</sup> in its UniProtKB entry, we can directly compute the number of distinct functions of the protein by classifying its annotated GO terms. GO is a pre-defined set of vocabulary organized in a hierarchical fashion. Thus, the similarity of GO terms can be objectively defined and computationally measured. In our earlier work,<sup>[18]</sup> we developed a procedure for clustering GO terms and identify moonlighting proteins and applied to the *E. coli* genome.

## Significance Statement

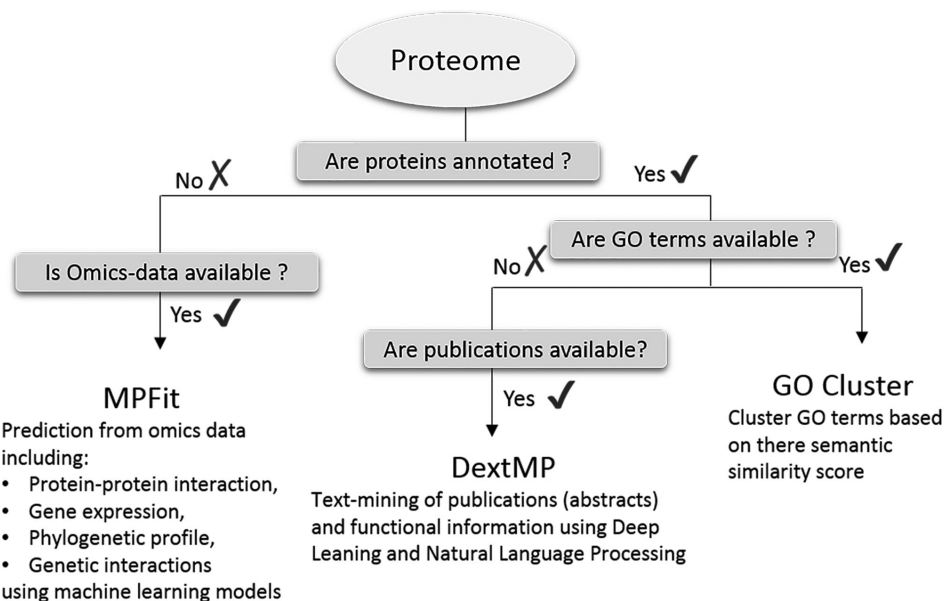
There is an increasing number of proteins that have been found to exhibit two distinct and independent functions called moonlighting proteins. Moonlighting proteins have been attracting attention recently because this concept requires us to update our fundamental understanding of protein functions. Moonlighting proteins also have strong implications in drug development and artificial protein design. In this article, we introduce our computational methods for systematic identification of moonlighting proteins in genomes. We applied one of the methods, which mines moonlighting proteins from literature, to three genomes and identified 13 new moonlighting proteins.

The second branch in Figure 1 is to handle proteins that have associated literature but no GO annotation. One can read literature related to candidate proteins, albeit a time-consuming effort if literature for many proteins needs to be examined. To overcome this issue, we have developed a machine learning based method, DextMP, which predicts if a protein moonlights or not based on text information, such as titles and abstracts of publications associated with the protein, or the functional information available in the UniProtKB database.<sup>[19]</sup> DextMP uses recent computational natural language processing techniques to encode the text information, which is later fed into several machine learning classifiers to identify potential moonlighting proteins.

Lastly, if several large-scale omics data for a protein are available, we can analyze the omics data to find characteristic patterns of moonlighting proteins in them. This is what the MPFit algorithm<sup>[20]</sup> is designed for, which corresponds to the left branch of Figure 1. MPFit is based on a simple and intuitive idea of moonlighting proteins. Since moonlighting proteins play a role in two (or more) different functions, they probably tend to interact with proteins from two (or more) different functional groups or pathways, and show correlated expression patterns and phylogenetic patterns<sup>[21]</sup> with proteins from two functional groups/pathways. Therefore, MPFit considers the number of functional clusters of interacting proteins in PPI, co-expression, and phylogenetic profile networks as features of a query protein and feeds it to a machine learning method (random forest) to make the prediction of moonlighting proteins.

It should be noted that these three methods are currently used for screening potential moonlighting proteins, and further manual verification, such as a careful reading of related literature, is needed to finalize a conclusion. This is because these methods do not examine the strict original definition of moonlighting proteins mentioned earlier and because the semantic distance of GO terms does not always capture the distinctiveness of functions well. For example, we occasionally encounter cases that a protein has GO terms that are distant on the GO hierarchy (and thus judged as potential moonlighting proteins) but these terms are somewhat related from a biological point of view. The latter problem is difficult to fix because it originated from the hierarchical graphical structure of current GO.

In this work, we ran DextMP on three genomes, *Arabidopsis thaliana*, *Caenorhabditis elegans*, and *Drosophila melanogaster*. In the original paper of DextMP,<sup>[19]</sup> the prediction accuracy was



**Figure 1.** Different methods for identifying moonlighting proteins in a genome. When a protein is annotated, clustering GO terms based on their similarity can identify multifunctional proteins. When a literature or functional description of the protein is available, the text mining tool, DextMP, can be used. The omics-data-based method MPFit is useful when a protein is not annotated but several other data, such as protein–protein interaction, expression profile, etc., is available.

benchmarked on known moonlighting proteins in *E. coli*, human, and mouse, stored in MoonProtDB. Since we considered that the accuracy observed was sufficiently high for further application (*F*-score, the harmonic mean of recall and precision of over 0.9), here we applied it to three genomes whose moonlighting proteins are not well studied.

After screening text information of proteins in the three genomes, we performed a two-step manual literature check. During the process, we identified a problem caused by “hub publications”, papers that are associated with several proteins. Generally, such papers comprise of large-scale genomics and proteomics experiment. Hub publications tend to cause false positives, because multiple proteins, thus multiple functions, are mentioned in the text. We removed hub publications to reduce false positives and thus to reduce the burden of downstream manual literature check steps. We identified 13 new moonlighting proteins, which we classified into two classes depending on the confidence level. Finally, four proteins in the high confidence level class are discussed individually.

## 2. Experimental Section

First, the analyzed genomes and the text information of proteins used for detecting moonlighting proteins were described. Then, the overall procedure used to identify moonlighting proteins in the genome was explained, and then the algorithm of DextMP was described.

### 2.1. Genome Dataset

DextMP was run on proteins in three genomes, *Arabidopsis thaliana*, *Caenorhabditis elegans*, and *Drosophila melanogaster*.

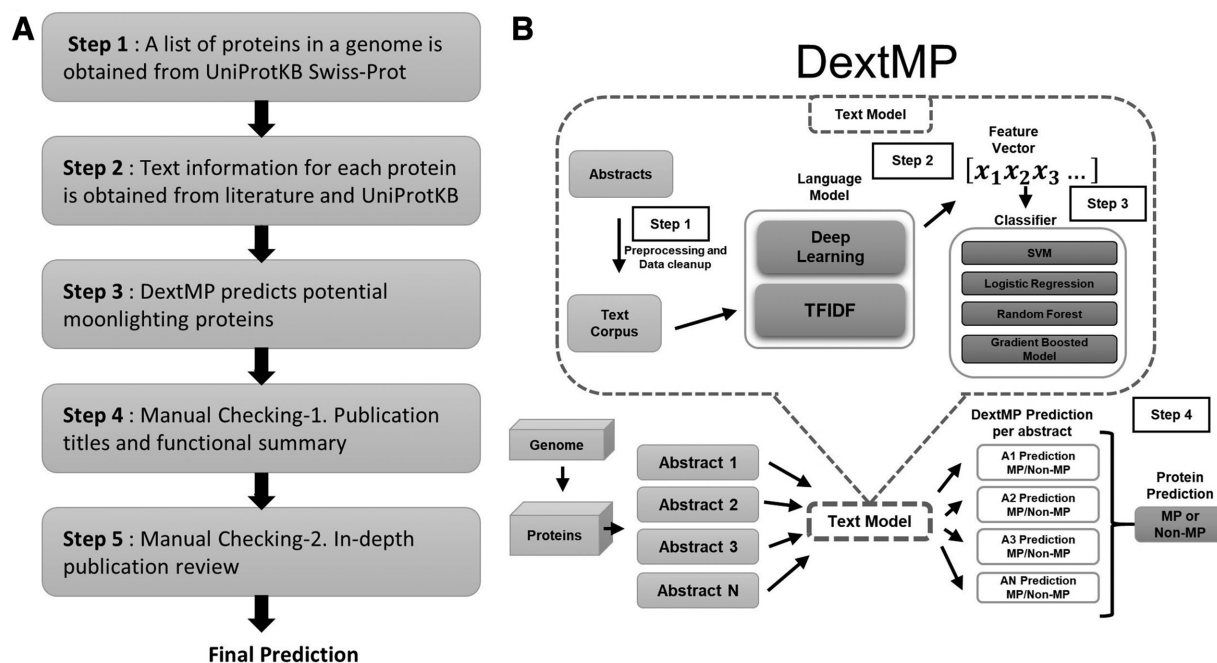
Three criteria were applied for choosing these genomes. First, model organisms were analyzed, because they were relatively well studied and had abundant publications in PubMed. Second, among popular model organisms, human, yeast, and *Xenopus laevis* were excluded, because they were analyzed in the original paper of DextMP. *Escherichia coli* and mouse genomes were also avoided, because moonlighting proteins of these two genomes were abundant in MoonProtDB and were used for the parameter training of DextMP.

### 2.2. Text Information of Proteins

For each protein, three types of textual information were extracted. First, a title of each publication of the protein, which was obtained from the list of “PUBLICATIONS” in its UniProtKB entry. Second, an abstract of each publication, which was extracted from the PubMed database (<https://www.ncbi.nlm.nih.gov/pubmed/>) using the PubMed ID of the publication as the key for the database search. Third, the functional description text of the protein, which was obtained from the function subsection in the “FUNCTION” section in its UniProtKB entry. The title and function description were downloaded from <https://www.uniprot.org/downloads>.

### 2.3. Overall Procedure of Identifying Moonlighting Proteins

The procedure consisted of five steps (Figure 2A): 1) Proteins in a genome were obtained from UniProtKB Swiss-Prot. 2) For each protein, three different types of text information, literature titles, and abstracts as well as function summary description from UniProtKB were obtained. Hub publications were omitted and



**Figure 2.** Procedure of identifying moonlighting proteins used in this work. A) Overall procedure. Proteins in a genome were obtained from UniProtKB Swiss-Prot. Three types of literature information: publication titles, publication abstract, and UniProtKB functional summary about the proteins were extracted. DextMP predicted if a protein is a potential moonlighting protein or not based on publication abstracts. Predicted proteins underwent a quick Manual Checking-1, and those which passed are checked again in Manual Checking-2 by careful reading of the literature to finalize the list of moonlighting proteins. B) The DextMP algorithm. There are four steps in the algorithm: 1) Each abstract is cleaned and processed, which involved removal of stop words, punctuation, and special symbols, followed by stemming and lemmatization. 2) Each of two language models, a deep learning model and TFIDF, converted the cleaned text into a feature vector. 3) The feature vector (representing one abstract) was predicted as 1 (moonlighting) or 0 (non-moonlighting). 4) A majority vote was applied to predictions made for the entire abstracts of a protein to predict if the protein is moonlighting or non-moonlighting.

consequently, proteins that only had hub publications that associated with more than three proteins were removed. 3) DextMP was used to predict if a protein was moonlighting or not from publication abstracts of the protein. 4) Predicted moonlighting proteins were manually examined by quickly checking publication titles and the functional description in UniProtKB (Manual Checking-1). Both text information can provide an indication if two different functions were associated with the protein. 5) Proteins that passed Step 4 underwent Manual Checking-2, which was an in-depth literature review of the protein. This was the final step where the two functions of the proteins were confirmed as independent from each other by reading the literature.

## 2.4. The DextMP Algorithm

DextMP took textual information of proteins to predict if a protein was moonlighting or not using machine learning methods. There were four steps in the DextMP algorithm (Figure 2B).

First, input text of a query protein underwent data clean-up. In the original work of DextMP,<sup>[19]</sup> three different types of text information were tested, which were publication titles, publication abstracts, and UniProtKB functional summary. Among the three input types, using publication abstracts showed the highest accuracy.<sup>[19]</sup> Hence, in this work, abstracts were used as the protein text information. Cleaning up of text data (abstracts)

involved removal of stop words, punctuations, and symbols. Next, stemming and lemmatization was done using the nltk package (a natural language analysis toolkit).

In the second step, the cleaned text (abstract) was converted into a *k*-dimensional feature vector using a statistical language model. Based on the accuracy reported in the original DextMP paper,<sup>[19]</sup> two best language models were used, which were term frequency inverse document frequency (TFIDF)<sup>[22]</sup> and a deep neural network named the paragraph vector.<sup>[23]</sup> to construct the feature vector. TFIDF is a vector that is computed from the number of counts of each word in the abstract relative to the frequency of words observed in the text corpus, which is a dictionary of words taken from all abstracts in a dataset. As the text corpus, a dataset of abstracts was used for 263 moonlighting proteins and 162 non-moonlighting proteins, which were collected in the original DextMP paper. The deep neural network learning language model mapped a text (an abstract) into a vector space using a neural network, which was trained in a way that semantically similar texts appeared closer in the vector space.<sup>[23]</sup> Thus, intuitively, a vector constructed by the deep neural network captured similarities of abstracts.

Subsequently (Step 3), each of four machine learning methods, linear regression (LR), support vector machine (SVM), random forest (RF), and the gradient boosted machine (GBM), took the input feature vector and classified it into moonlighting or non-moonlighting. The prediction was binary, and each abstract



**Table 1.** The number of proteins in each genome selected by each step of the procedure.

Genome	After removing hub publications	Predicted as moonlighting by DextMP	Passed Manual Checking-1	Passed Manual Checking-2
<i>Arabidopsis thaliana</i>	7045	1,917	23	7
<i>Caenorhabditis elegans</i>	1600	1,193	16	3
<i>Drosophila melanogaster</i>	2663	2,86	19	2

associated with a query protein was predicted as 1 (moonlighting) or 0 (non-moonlighting).

In the last step, the prediction made for each abstract of a protein was summarized by majority vote to make the final prediction. Combinations of two language models (TFIDF or the deep learning model) and four machine learning classifiers (LR, SVM, RF, or GBM) resulted in eight final predictions for a protein. If a protein was predicted to be a moonlighting protein by at least one of the language models—classifier combinations, the protein was considered a candidate for moonlighting and passed to the manual checking steps (Figure 2A).

The parameters of the language models and the machine learning methods were trained on the same dataset that was used in the original DextMP paper.<sup>[19]</sup> The accuracy of DextMP on the training dataset ranged from 0.716 to 0.936 for different combinations of language models—classifiers, which were comparable to the values reported in the original paper. The program can be downloaded from <http://kiharalab.org/DextMP>.

### 3. Results

#### 3.1. Identifying Moonlighting Proteins

We ran our procedure (Fig. 2) to identify moonlighting proteins on the three genomes. **Table 1** shows the number of proteins that were selected at each step of the procedure. In the *A. thaliana*, *C. elegans*, and *D. melanogaster* genome, 7,045, 1,600, and 2,663 proteins, respectively, had at least one publication after removing hub publications, which appear as reference for more than three proteins. Among them, 1,917, 1,193, and 286 proteins, respectively, were predicted as moonlighting proteins by DextMP. Manual Checking-1 reduced the potential moonlighting proteins significantly, to 1.2, 1.3, and 6.6% for the three genomes. In this step, we only checked titles of publications and UniProt function summary of proteins, because paper abstracts were considered as input data of DextMP in the previous step. Manual Checking-2 which involves careful and thorough reading, finally predicted 13 new moonlighting proteins. A summary of the proteins is provided in **Table 2**.

In general, there were two reasons for a protein that passed Manual Checking-1 was not judged as a moonlighting protein in Manual Checking-2 when we read the abstract and the main text of papers. The first case is that papers made it clear that the protein was not moonlighting. For example, matrix metalloproteinase-2 in *Drosophila* (UniProtKB ID: Q8MPP3) has several publications indicating multiple non-related functions such as tissues remodeling, motor neurons contraction, and reepithelization during wound healing. However, when we investigated the mechanism of these functions, we found that in all the

biological processes mentioned, the protein performs the same proteinase activity. The second case is that there is not enough information available to conclude that a protein is a moonlighting protein. Tyrosine-protein kinase csk-1 (C-terminal Src kinase) (UniProtKB ID: G5ECJ6) was such an example. csk-1 is known to regulate Src family tyrosine kinases (SFKs). In *C. elegans*, two SFK's, src-1 and src-2, are identified, and it has been shown that csk-1 specifically targets the C-terminal tyrosine of both src-1 and src-2, negatively regulating their activities.<sup>[24]</sup>{Hirose, 2003 #355} We found a paper that showed csk-1 is important for pharyngeal muscle organization, independent of src-1 and src-2.<sup>[25]</sup> This piqued our interest; however, we found that src-2 is important for larval and pharynx development, thus, csk-1 affects the pharynx development both with and without the SFK's involvement. Further, it is suggested by the authors that csk-1 might interact with another unknown protein to control pharynx development by phosphorylating a tyrosine of the protein. Thus, we concluded that csk-1 probably only has the kinase function and might not perform a second function in pharyngeal muscle organization.

#### 3.2. Case Studies

In **Table 1**, we classified the predicted proteins into two classes based on the confidence of independence of two functions of the proteins. For class 1 proteins there is a clear indication in literature that the two functions are independent of each other or that the functions are performed in different locations in the organism. Proteins which seemingly have two separate functions, but their independence is not well established by current knowledge are categorized as Class 2. In the table, we also show the number of domains defined in the Pfam database,<sup>[26]</sup> as protein multifunctionality attributed to either gene fusion event or presence of multiple domains is generally not considered as moonlighting in the original definition. Below, we discuss the four Class 1 potential moonlighting proteins.

#### 3.3. Chloroplastic Leucine Aminopeptidase 2

The first example is leucine aminopeptidase 2 (LAP2) from *Arabidopsis* (UniProtKB ID: Q944P7). It is a di-zinc metallopeptidase that catalyzes the cleavage of amino acids from N-terminal of various peptides. In the paper by Scanton et al., the peptidase activity of LAP2 was demonstrated on a model substrate, leucine-amino methyl coumarin.<sup>[27]</sup> In the paper, LAP2 has been shown to have chaperone activity as well. Chaperones are proteins that assist other proteins in folding and unfolding. The authors discovered that LAP2 possess chaperone activity by observing that

LAP2 prevented the thermal inactivation of two tested proteins Luc and Ndel. It was further shown that the chaperone function of LAP2 was independent of its peptidase function by mutating the amino acids responsible for peptidase function.<sup>[27]</sup> This is a relatively easy example to detect from literature because the title and the abstract of this paper used the word “moonlighting”.

### 3.4. Actin-Depolymerizing Factor 9

The second protein is actin-depolymerizing factor 9 (ADF9) in *A. thaliana* (UniProtKB ID: O49604). It stabilizes actin filaments and acts as an antagonist to other ADF's. In the presence of ADF9, the acting filaments organize themselves into actin bundles, which is similar to other actin bundling protein actions. This function has been confirmed in vitro as well as in vivo.<sup>[28]</sup> ADF9 is also important for the expression of flowering locus C (FLC) gene, which is responsible for flowering, indicating that ADF9 is a potential moonlighting protein. The *adf9* mutation decreased the level of histone H3 at multiple sites of FLC promoter region, indicating that ADF9 helps in maintaining the chromatin remodeling machinery intact, which regulates the FLC expression.<sup>[29]</sup>

### 3.5. Dihydrofolate Reductase

Dihydrofolate reductase (DHFR) (UniProtKB ID: P17719) is an important enzyme in the folate biosynthesis pathway, where it synthesizes 5,6,7,8-tetrahydrofolate from 7,8-dihydrofolate.<sup>[30]</sup> DHFR is already a known moonlighting protein in human, where aside from its enzymatic activity, it also possesses the ability to bind RNA. DHFR in human binds to DHFR mRNA, thus regulating its own synthesis.<sup>[31]</sup> The DHFR of *Drosophila* is a moonlighting protein as well, as it also plays a role in cell survival by interacting with another protein, known as vestigial protein, and controlling gene expression.<sup>[32]</sup> Vestigial protein (vg) regulates the formation of wings by interacting with nuclear regulatory proteins and controlling gene expression in the wing region. It has been observed that vg regulates DHFR expression at the D/V boundary in the wing disc of *Drosophila*. Also, decrease in DHFR (and vg) leads to caspase mediated cell death and wing margin defects.<sup>[32]</sup> Thus, the second function of DHFR of *Drosophila* is different from that of human DHFR. As we see in this example, it is not uncommon for a homologous protein of a moonlighting protein to either not have a secondary function<sup>[33]</sup> or has a different secondary function.<sup>[34]</sup>

### 3.6. Twinkle Homolog Protein

The last Class 1 moonlighting protein, DNA helicase (UniProtKB ID: B5 × 582), is a protein that can unwind the double-stranded DNA helix into separate strands and opens the DNA to be used as a template for DNA replication. On the other hand, DNA primase is an enzyme that catalyzes the synthesis of a small single stranded RNA that helps in DNA replication. Generally, these two functions are performed by two separate enzymes. The T7

bacteriophage gp4 proteins, however, is a multifunctional protein that has both helicase and primase activity.<sup>[35]</sup> The twinkle homolog protein is homologous to the gp4 protein of T7 bacteriophage. Such homologs are present in several eukaryotes where they function as only DNA helicases, losing their DNA primase activity. The twinkle protein in *A. thaliana* is found to possess both DNA helicase as well as primase activity, making it a dual-functional protein. It is present in chloroplast and mitochondria where its primase functions to produce RNA primers, which may help in organelle DNA replication.<sup>[35]</sup>

The two functions of this protein are performed by different domains as shown in Table 1 the primase function is carried out by the toprim domain (Pfam ID: PF13662, Toprim\_4) while the helicase activity is performed by DnaB-like helicase C-terminal domain (Pfam ID: PF03796, DNB-C). Note that the two-domain structure may disqualify this protein from being as moonlighting proteins by the original definition because it considers only cases where bi-functionality is not due to multiple domains or gene fusion events.

### 3.7. Class 2 Moonlighting Proteins

Table 1 includes nine Class 2 moonlighting proteins. For Class 2 proteins, two functions are described in the literature but due to the lack of experimental evidence, it was unclear if one of the functions is not an outcome of the other function. Class 2 category also includes cases that one of the functions is assumed from sequence similarity to a homologous protein. Since homologous proteins do not always share moonlighting function, the assumed functions need to be verified by experiments.

## 4. Discussion

Moonlighting proteins are shedding new light on functional studies of genomes and proteomes. The increasing number of identified moonlighting proteins suggests that multiplicity of functions of proteins would always need to be considered for functional studies. Information for the multiple functions associated with a protein is listed in the UniProt, but is only indicated in the functional description. As moonlighting in proteins is a fairly new concept, the database has not provided a specific label that indicates moonlighting, or more generally, bi-functionality, which makes a systematic study difficult.

The most accurate approach to identify moonlighting proteins is to manually read the published literature, that is, to search for proteins that have been experimentally confirmed to perform two or more functions. However, going through a huge amount of publications is a daunting and prohibitively time-consuming task. Our group has previously developed a text-mining tool, DextMP, which takes text information from publications or functional descriptions in UniProtKB and predicts if a protein moonlights or not. DextMP can computationally screen literature and database entries of thousands of proteins in a genome and provides a short list of potential moonlighting proteins, significantly reducing the load of users in checking the literature. In this work, we performed genome-wide moonlighting protein identification

**Table 2.** List of identified moonlighting proteins.

Name	UniProtKB ID	Function 1 (F1)	Function 2 (F2)	Reference for F1	Reference for F2	Confidence class	Number of Pfam domains
<i>Arabidopsis thaliana</i>							
Leucine aminopeptidase 2, chloroplastic	Q944P7	Molecular chaperones	Leucine aminopeptidase activity, role in insect defense	[27]	[27]	1	1
BTB/POZ and TAZ domain-containing protein 2	Q94BN0	component of the TAC1-mediated telomerase activation pathway	mediating diverse hormone, stress, and metabolic responses	[36]	[37]	2	2
Chromophore lyase CRL, chloroplastic	Q9F146	Required for plastid division, and involved in cell differentiation and regulation of the cell division plane	Confers sensitivity to cabbage leaf curl virus, probably by hindering its movement	[38]	[39]	2	1
Twinkle homolog protein	B5 × 582	DNA helicase	DNA primase	[35]	[35]	1	2
Alkaline/neutral invertase C, mitochondrial	B9DFA8	Mitochondrial invertase that cleaves sucrose into glucose and fructose	Regulation of aerial tissue development	[40]	[41]	2	1
Actin-depolymerizing factor 9	O49606	Stabilize and cross-link actin filaments	Controls expression of Flowering Locus C gene via controlling chromatin remodeling	[28]	[29]	1	1
NEDD8-activating enzyme E1 regulatory subunit AXR1	P42744	Regulatory subunit ECR1-AXR1 E1 enzyme	Regulates the chromosomal localization of meiotic recombination by crossovers and subsequent synapsis, probably through the activation of a CRL4 complex	[42]	[43]	2	1
<i>Drosophila melanogaster</i>							
Dihydrofolate reductase	P17719	Interact with vestigial, this interaction may be important for cell proliferation and survival	Dihydrofolate reductase activity	[32]	[30]	1	1
Lon protease homolog, mitochondrial	Q7KUT2	ATP dependent serine protease	Chaperone function in assembly of inner membrane protein complexes	[44]	By similarity [45]	2	3

(Continued)

**Table 2.** Continued.

Name	UniProtKB ID	Function 1 (F1)	Function 2 (F2)	Reference for F1	Reference for F2	Confidence class	Number of Pfam domains
<i>Caenorhabditis elegans</i> Exchange factor for Arf-6	G5EET6	guanine nucleotide exchange factor for ARF6	Limit microtubule growth independent of arf-6, inhibit axon regrowth	By similarity <sup>[46]</sup>	[47]	2	2
Matrix metalloproteinase-B	O44836	embryogenesis/adult development	pathogen resistance	[48]	[48]	2	1
Caveolin-2	Q18879	Scaffolding protein within caveolar membrane	uptake of lipids and proteins in intestinal cells	By similarity <sup>[49]</sup>	[50]	2	1
Mitochondrial 2-oxoglutarate/malate carrier protein	P90992	Transfer alpha ketoglutarate across inner mitochondrial membrane	Control apoptosis through LIN-35/RB-like protein pathway	By similarity <sup>[51]</sup>	[52]	2	1

for three genomes. From the short list provided by DextMP, we applied a two-step manual literature and database check to find promising moonlighting proteins. The first manual screening, Manual Checking-1, i.e., checking literature titles and UniProtKB functional summary, was introduced for efficiency, and indeed significantly contributed by speeding up the entire manual check process. On the other hand, it is highly possible that some genuine moonlighting proteins were missed by this step. In practice, there is a tradeoff between the time requirement and finding more moonlighting proteins by a careful and thorough reading of literature. Manual Checking-2 is a thorough analysis of the publications related to proteins. Specifically, we looked for evidence where inhibition of one function does not affect the second function and vice versa, confirming that both functions are independent. Upon further improvement of the accuracy of DextMP, we aim to substantially reduce the manual post-processing step; possibly, even removing the manual steps entirely. Below, we discuss several directions for improvement of DextMP.

While running DextMP, we discovered that hub publications, papers that link to several proteins, confuse the program to classify them as moonlighting proteins. A preprocessing step, where such papers are removed, is crucial to reduce the number of false positives in the predictions.

We found that another source of false positives originated from different levels of function descriptions in literature. For example, there are often cases where protein functions are discussed at both molecular and biological levels. At a molecular level, a protein's interacting partners, biological pathways the protein belongs to, or active site residues are described whereas biological level descriptions include how the protein affects at a cell or organism level, such as the development, proliferation, and embryogenesis. Currently, DextMP cannot distinguish these two types of functions, and therefore whenever both levels of information are written, the algorithm identifies it as two independent functions and classifies the publication as containing moonlighting protein information. This was observed during manual analyses of DextMP predictions. Being able to identify the two classes of functions mentioned in the paper will greatly improve the specificity of the model.

As shown in Figure 2, DextMP makes a prediction for an individual publication associated with the protein separately, which is then combined by a majority vote to make a final prediction. Therefore, a moonlighting protein will be missed if only one function is mentioned in each individual paper. Practically, this seemed not to be a large problem as usually a newer paper reporting novel secondary functions mention the original function of a protein in its abstract. To be able to consider the all papers for a protein together, technically we will need to introduce a way to judge similarity or dissimilarity of mentioned functions (i.e. different, thus potentially moonlighting function) across papers, which is an interesting technical challenge.

Finally, analyzing the whole papers instead of simply abstracts will provide more information and will contribute to making more accurate predictions, so long as useful information for classification can be effectively extracted from large text information.

Natural language processing (NLP) techniques are a fast-growing area in artificial intelligence research. By introducing new techniques in NLP, we hope to further improve DextMP and



contribute in deciphering complex functional interplay of proteins in the cell.

## Acknowledgements

The authors are thankful to Md. Mansurul Bhuiyan for his technical help in running DextMP. The authors are also thankful to Samarth Mathur, Myson C. Burch, and Lyman Monroe for proofreading the manuscript. This study is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the Army Research Office (ARO) under cooperative Agreement Number W911NF-17-2-0105. D.K. also acknowledges support from the National Institute of General Medical Sciences of the National Institutes of Health (R01GM123055) and the National Science Foundation (DMS1614777).

## Conflict of Interest

The authors declare no conflict of interest.

## Keywords

bifunctional proteins, functional genomics, moonlighting proteins, protein function annotation, text mining

Received: May 28, 2018

Revised: August 13, 2018

Published online: October 10, 2018

- [1] a) T. Hawkins, M. Chitale, S. Luban, D. Kihara, *Proteins* **2009**, *74*, 566; b) I. K. Khan, Q. Wei, M. Chitale, D. Kihara, *Bioinformatics* **2015**, *31*, 271; c) M. Chitale, T. Hawkins, C. Park, D. Kihara, *Bioinformatics* **2009**, *25*, 1739; Q. Wei, J. McGraw, I. Khan, D. Kihara, *Methods Mol. Biol.* **2017**, *1611*, 1; d) Y. Jiang, T. R. Oron, W. T. Clark, A. R. Bankapur, D. D'Andrea, R. Lepore, C. S. Funk, I. Kahanda, K. M. Verspoor, A. Ben-Hur, C. E. Koo da, D. Penfold-Brown, D. Shasha, N. Youngs, R. Bonneau, A. Lin, S. M. Sahraeian, P. L. Martelli, G. Profiti, R. Casadio, R. Cao, Z. Zhong, J. Cheng, A. Altenhoff, N. Skunca, C. Dessimoz, T. Dogan, K. Hakala, S. Kaewphan, F. Mehryary, T. Salakoski, F. Ginter, H. Fang, B. Smithers, M. Oates, J. Gough, P. Toronen, P. Koskinen, L. Holm, C. T. Chen, W. L. Hsu, K. Bryson, D. Cozzetto, F. Minneci, D. T. Jones, S. Chapman, D. Bkc, I. K. Khan, D. Kihara, D. Ofer, N. Rappoport, A. Stern, E. Cibrian-Uhalte, P. Denny, R. E. Foulger, R. Hieta, D. Legge, R. C. Lovering, M. Magrane, A. N. Melidoni, P. Mutowo-Muullenet, K. Pichler, A. Shypitsyna, B. Li, P. Zakeri, S. ElShal, L. C. Tranchevent, S. Das, N. L. Dawson, D. Lee, J. G. Lees, I. Sillitoe, P. Bhat, T. Nepusz, A. E. Romero, R. Sasidharan, H. Yang, A. Paccanaro, J. Gillis, A. E. Sedeno-Cortes, P. Pavlidis, S. Feng, J. M. Cejuela, T. Goldberg, T. Hamp, L. Richter, A. Salamov, T. Gabaldon, M. Marcet-Houben, F. Supek, Q. Gong, W. Ning, Y. Zhou, W. Tian, M. Falda, P. Fontana, E. Lavezzo, S. Toppo, C. Ferrari, M. Giollo, D. Piovesan, S. C. Tosatto, A. Del Pozo, J. M. Fernandez, P. Maietta, A. Valencia, M. L. Tress, A. Benso, S. Di Carlo, G. Politano, A. Savino, H. U. Rehman, M. Re, M. Mesiti, G. Valentini, J. W. Bargsten, A. D. van Dijk, B. Gemovic, S. Glisic, V. Perovic, V. Veljkovic, N. Veljkovic, E. S. D. C. Almeida, R. Z. Vencio, M. Sharan, J. Vogel, L. Kansakar, S. Zhang, S. Vucetic, Z. Wang, M. J. Sternberg, M. N. Wass, R. P. Huntley, M. J. Martin, C. O'Donovan, P. N. Robinson, Y. Moreau, A. Tramontano, P. C. Babbitt, S. E. Brenner, M. Linial, C. A. Orengo, B. Rost, C. S. Greene, S. D. Mooney, I. Friedberg, P. Radivojac, *Genome Biol.* **2016**, *17*, 184.
- [2] J. Piatigorsky, W. E. O'Brien, B. L. Norman, K. Kalumuck, G. J. Wistow, T. Borrás, J. M. Nickerson, E. F. Wawrousek, *Proc. Natl. Acad. Sci. U. S. A.* **1988**, *85*, 3479.
- [3] C. Pandya, J. D. Farelli, D. Dunaway-Mariano, K. N. Allen, *J. Biol. Chem.* **2014**, *289*, 30229.
- [4] C. J. Jeffery, *Trends Biochem. Sci.* **1999**, *24*, 8.
- [5] E. Chu, D. M. Koeller, J. L. Casey, J. C. Drake, B. A. Chabner, P. C. Elwood, S. Zinn, C. J. Allegra, *Proc. Natl. Acad. Sci. U. S. A.* **1991**, *88*, 8977.
- [6] M. J. Stutts, C. M. Canessa, J. C. Olsen, M. Hamrick, J. A. Cohn, B. C. Rossier, R. C. Boucher, *Science* **1995**, *269*, 847.
- [7] a) C. J. Jeffery, *IUBMB Life* **2011**, *63*, 489; b) G. Sriram, J. A. Martinez, E. R. McCabe, J. C. Liao, K. M. Dipple, *Am. J. Hum. Genet.* **2005**, *76*, 911.
- [8] C. J. Jeffery, *Front. Genet.* **2015**, *6*, 211.
- [9] C. Chen, S. Zabad, H. Liu, W. Wang, C. Jeffery, *Nucleic Acids Res.* **2018**, *46*, D640.
- [10] L. Franco-Serrano, S. Hernandez, A. Calvo, M. A. Severi, G. Ferragut, J. Perez-Pons, J. Pinol, O. Pich, A. Mozo-Villarias, I. Amela, E. Querol, J. Cedano, *Nucleic Acids Res.* **2018**, *46*, D645.
- [11] I. K. Khan, D. Kihara, *Biochem. Soc. Trans.* **2014**, *42*, 1780.
- [12] a) A. Gomez, N. Domedel, J. Cedano, J. Pinol, E. Querol, *Bioinformatics* **2003**, *19*, 895; b) S. Hernandez, L. Franco, A. Calvo, G. Ferragut, A. Hermoso, I. Amela, A. Gomez, E. Querol, J. Cedano, *Front. Bioeng. Biotechnol.* **2015**, *3*, 90; c) I. Khan, M. Chitale, C. Rayon, D. Kihara, *BMC Proc.* **2012**, *6*(Suppl 7), S5.
- [13] a) S. Hernandez, I. Amela, J. Cedano, J. Pinol, J. Perez-Pons, A. Mozo-Villarias, E. Querol, *J. Proteomics Bioinform.* **2012**, *5*, 262; b) P. Tompa, C. Szasz, L. Buday, *Trends Biochem. Sci.* **2005**, *30*, 484; c) H. J. Dyson, *Q. Rev. Biophys.* **2011**, *44*, 467.
- [14] a) A. Gomez, S. Hernandez, I. Amela, J. Pinol, J. Cedano, E. Querol, *Mol. Biosyst.* **2011**, *7*, 2379; b) Y. Pritykin, D. Ghersi, M. Singh, *PLoS Comput. Biol.* **2015**, *11*, e1004467; c) C. E. Chapple, B. Robisson, L. Spinelli, C. Guien, E. Becker, C. Brun, *Nat. Commun.* **2015**, *6*, 7412.
- [15] L. Cheng, K. S. Leung, *Bioinformatics* **2018**. <https://doi.org/10.1093/bioinformatics/bty399>
- [16] S. Pundir, M. J. Martin, C. O'Donovan, *Methods Mol. Biol.* **2017**, *1558*, 41.
- [17] G. O. Consortium, *Nucleic Acids Res.* **2015**, *43*, D1049.
- [18] I. Khan, Y. Chen, T. Dong, X. Hong, R. Takeuchi, H. Mori, D. Kihara, *Biol. Direct* **2014**, *9*, 30.
- [19] I. K. Khan, M. Bhuiyan, D. Kihara, *Bioinformatics* **2017**, *33*, i83.
- [20] a) I. Khan, J. McGraw, D. Kihara, *Methods Mol. Biol.* **2017**, *1611*, 45; b) I. K. Khan, D. Kihara, *Bioinformatics* **2016**, *32*, 2281.
- [21] M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, T. O. Yeates, *Proc. Natl. Acad. Sci. U. S. A.* **1999**, *96*, 4285.
- [22] C. D. Manning, P. Raghavan, H. Schütze, *Introduction to Information Retrieval*, Cambridge University Press, Cambridge, UK **2008**.
- [23] Q. Le, T. Mikolov, *Proc. Machine Learn. Res.*, **2014**, *32*, 1188.
- [24] T. Hirose, M. Koga, Y. Ohshima, M. Okada, *FEBS Lett.* **2003**, *534*, 133.
- [25] N. Takata, B. Itoh, K. Misaki, T. Hirose, S. Yonemura, M. Okada, *Genes Cells* **2009**, *14*, 381.
- [26] R. D. Finn, P. Coghill, R. Y. Eberhardt, S. R. Eddy, J. Mistry, A. L. Mitchell, S. C. Potter, M. Punta, M. Qureshi, A. Sangrador-Vegas, G. A. Salazar, J. Tate, A. Bateman, *Nucleic Acids Res.* **2016**, *44*, D279.
- [27] M. A. Scranton, A. Yee, S. Y. Park, L. L. Walling, *J. Biol. Chem.* **2012**, *287*, 18408.
- [28] S. Tholl, F. Moreau, C. Hoffmann, K. Arumugam, M. Dieterle, D. Moes, K. Neumann, A. Steinmetz, C. Thomas, *FEBS Lett.* **2011**, *585*, 1821.
- [29] B. Burgos-Rivera, D. R. Ruzicka, R. B. Deal, E. C. McKinney, L. King-Reid, R. B. Meagher, *Plant Mol. Biol.* **2008**, *68*, 619.

- [30] H. Hao, M. G. Tyshenko, V. K. Walker, *J. Biol. Chem.* **1994**, *269*, 15179.
- [31] E. Chu, C. H. Takimoto, D. Voeller, J. L. Grem, C. J. Allegra, *Biochemistry* **1993**, *32*, 4756.
- [32] R. Delanoue, K. Legent, N. Godefroy, D. Flagiello, A. Dutriaux, P. Vaudin, J. L. Becker, J. Silber, *Cell Death Differ.* **2004**, *11*, 110.
- [33] P. Ozimek, P. Kotter, M. Veenhuis, I. J. van der Klei, *FEBS Lett.* **2006**, *580*, 46.
- [34] a) S. Banerjee, A. K. Nandyala, P. Raviprasad, N. Ahmed, S. E. Hasnain, *J. Bacteriol.* **2007**, *189*, 4046; b) X. J. Chen, X. Wang, B. A. Kaufman, R. A. Butow, *Science* **2005**, *307*, 714; c) Y. Tang, J. R. Guest, *Microbiology* **1999**, *145*, 3069.
- [35] J. Diray-Arce, B. Liu, J. D. Cupp, T. Hunt, B. L. Nielsen, *BMC Plant Biol.* **2013**, *13*, 36.
- [36] S. Ren, K. K. Mandadi, A. L. Boedeker, K. S. Rathore, T. D. McKnight, *Plant Cell* **2007**, *19*, 23.
- [37] K. K. Mandadi, A. Misra, S. Ren, T. D. McKnight, *Plant Physiol.* **2009**, *150*, 1930.
- [38] T. Asano, Y. Yoshioka, S. Kurei, W. Sakamoto, Y. Machida, *Plant J.* **2004**, *38*, 448.
- [39] D. L. Trejo-Saavedra, J. P. Vielle-Calzada, R. F. Rivera-Bustamante, *Virol. J.* **2009**, *6*, 169.
- [40] X. Ji, W. Van den Ende, A. Van Laere, S. Cheng, J. Bennett, *J. Mol. Evol.* **2005**, *60*, 615.
- [41] M. L. Martin, L. Lechner, E. J. Zabaleta, G. L. Salerno, *Planta* **2013**, *237*, 813.
- [42] S. K. Hotton, R. A. Eigenheer, M. F. Castro, M. Bostick, J. Callis, *Plant Mol. Biol.* **2011**, *75*, 515.
- [43] M. T. Jahns, D. Vezon, A. Chambon, L. Pereira, M. Falque, O. C. Martin, L. Chelysheva, M. Grelon, *PLoS Biol.* **2014**, *12*, e1001930.
- [44] Y. Matsushima, Y. Goto, L. S. Kaguni, *PNAS* **2010**, *107*, 18410.
- [45] S. Goto-Yamada, S. Mano, C. Nakamori, M. Kondo, R. Yamawaki, A. Kato, M. Nishimura, *Plant Cell Physiol.* **2014**, *55*, 482.
- [46] J. E. Casanova, *Traffic* **2007**, *8*, 1476.
- [47] S. M. O'Rourke, S. N. Christensen, B. Bowerman, *Nat. Cell Biol.* **2010**, *12*, 1235.
- [48] B. Altincicek, M. Fischer, K. Luersen, M. Boll, U. Wenzel, A. Vilcinskis, *Dev. Comp. Immunol.* **2010**, *34*, 1160.
- [49] T. M. Williams, M. P. Lisanti, *Genome Biol.* **2004**, *5*, 214.
- [50] S. Parker, D. S. Walker, S. Ly, H. A. Baylis, *Mol. Biol. Cell* **2009**, *20*, 1763.
- [51] V. Iacobazzi, F. Palmieri, M. J. Runswick, J. E. Walker, *DNA Sequence* **1992**, *3*, 79.
- [52] M. Gallo, D. Park, D. S. Luciani, K. Kida, F. Palmieri, O. E. Blacque, J. D. Johnson, D. L. Riddle, *PLoS One* **2011**, *6*, e17827.