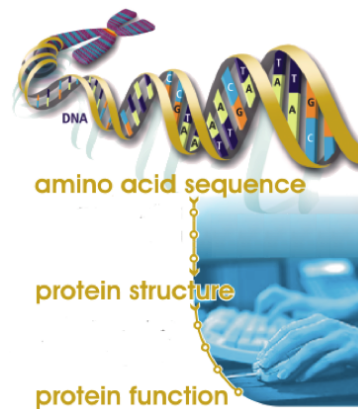




**SAPIENZA**  
UNIVERSITÀ DI ROMA

**DOTTORATO DI RICERCA IN BIOCHIMICA**  
**CICLO XXIV (A.A. 2008-2011)**

**From Linear Motif Discovery  
to Protein Function Detection**



**Docente guida**  
Prof.ssa Anna Tramontano

**Coordinatore**  
Prof. Paolo Sarti

*Dottorando*  
**Ahmed Sayadi**

**Dicembre 2011**





**SAPIENZA**  
UNIVERSITÀ DI ROMA

**DOTTORATO DI RICERCA IN BIOCHIMICA**  
**CICLO XXIV (A.A. 2008-2011)**

**From Linear Motif Discovery**  
**to Protein Function Detection**

**Docente guida**  
Prof.ssa Anna Tramontano

**Coordinatore**  
Prof. Paolo Sarti

*Dottorando*  
**Ahmed Sayadi**

**Dicembre 2011**

*To the exceptions  
and the laws governing them*



## **Acknowledgements**

It is a pleasure to thank the many people who made this thesis possible.

In the first place I would like to record my gratitude to my Ph.D. supervisor, Prof. Anna Tramontano for her help, her inspiration, her enthusiasm, her great efforts in explaining things clearly to me, and also for the great opportunity that she gave me, to learn from her and to enrich my knowledge as a student, a researcher and a scientist. I'm indebted to her more than she knows.

I would like to thank all the many people who have taught me bioinformatics: especially my master thesis supervisors (Toby Gibson, Gilles Travé) and my Ph.D. co-supervisor (Allegra Via) for their kind assistance, writing letters, giving wise advice and helping me to take the right decision.

I am grateful to all the colleagues at the Department of Biochemical Sciences "A. Rossi Fanelli" of the "Sapienza" University of Rome, and also to all the present and past members of the biocomputing group. A special thanks goes to Angela Tilia and Prof. Paolo Sarti for their precious help.

I gratefully acknowledge my deep sense of gratitude to all the people I worked and collaborated with, during my Ph.D. thesis: Allegra, Fabrizio, Paolo and Marco, for their valuable advice in scientific discussions, supervision and friendship. I hope to maintain our collaboration in the future.

I am grateful to the King Abdullah University of Science and Technology (KAUST) that supported me and funded the projects I have been working on during my Ph.D. period.

Words fail me to express my appreciation to my wife Wiem who dedicated her love and time to support me.

Last but not least, I wish to thank my friends Carmen and Daniel, my brothers Amine and Akram, my sister Zahra and most importantly my parents, Fredj and Najet. They were always present beside me and they offered me all the encouragement and help that I needed.

# Index

<b>Acknowledgements</b>	<b>3</b>
<b>Index</b>	<b>5</b>
<b>1. Introduction</b>	<b>7</b>
1.1. <i>Protein-Protein interaction modes</i>	8
1.2. <i>Short Linear Motifs</i>	9
1.2.1. What are Short Linear Motifs (SLiMs)?	9
1.2.1.1. SLiMs biological features	11
1.2.2. Why are we interested in Linear Motifs?	12
1.3. <i>Linear motifs discovery</i>	13
1.4. <i>Aim and contributions of the study</i>	14
<b>2. Methods aimed at identifying new instances of known LMs</b>	<b>16</b>
2.1. <i>ELM: The Eukaryote linear motif resource</i>	17
2.1.1. Results	18
2.1.1.1. mTOR related motifs	19
<b>3. <i>de novo</i> Linear motif prediction</b>	<b>26</b>
3.1. <i>MoDiPath: Motif Discovery in Pathways</i>	27
3.1.1. Results	27
3.1.1.1. Rediscovered motif: "SKL\$"	32
3.1.1.2. Newly discovered motif: "[FL].L.C..Y..A"	36
<b>4. Computational approach to the study of the g14-3-3 interactor network</b>	<b>39</b>
4.1. <i>g14-3-3 interactor network</i>	40
4.1.1. Results	41
4.1.1.1. Sequence analysis	41
4.1.1.2. Statistical significance estimation	41
4.1.1.3. Protein sequence alignments	43
4.1.1.4. Disorder prediction	44
4.1.1.5. Structure prediction	45

<b>5. Conclusions and outlook</b>	<b>47</b>
<b>6. Materials and Methods</b>	<b>49</b>
6.1. <i>CD-HIT</i>	49
6.2. <i>SlimFinder</i>	49
6.3. <i>CompariMotif</i>	50
6.4. <i>Conservation Score</i>	51
6.5. <i>IUPred</i>	51
6.6. <i>TwoSampleLogo</i>	51
6.7. <i>HHPred</i>	52
6.8. <i>QMEAN</i>	53
6.9. <i>NetSurfP</i>	53
6.10. <i>POPS</i>	54
<b>7. References</b>	<b>55</b>
<b>8. Publications</b>	<b>61</b>
<i>Paper I: ELM: the status of the 2010 eukaryotic linear motif resource</i>	62
<i>Paper II: Exploiting Biological and Biochemical Information</i>	77
<i>Paper III: The interaction network of the 14-3-3 protein in the ancient protozoan parasite Giarida duodenolis</i>	89

## 1. Introduction

Bioinformatics is the field of science in which biology, computer science, and information technology merge to form a single discipline.

The initial interest in bioinformatics started when Sanger discovered the method to sequence proteins (Sanger and Coulson, 1975). The vast amount of biological data that has become available after this discovery made the need to organize, analyze and store them pressing.

The first sequence database was created within a short period after the first protein sequence was made available in 1956 (Insulin protein sequence with 51 residues) (Stretton, 2002), nowadays the manually created SWISS-PROT database, which was created in 1986, (Bairoch and Boeckmann, 1994) counts more than 530.000 protein sequences (release November 2011) (<http://web.expasy.org/docs/relnotes/relstat.html>).

The availability of the entire genome sequences for several organisms and the exponential growth in computing power during the last few years have expanded the research focus from the study of a single molecule, gene, protein or small complex to the exhaustive exploration of molecular interactions and biological processes at the level of whole organisms. Proteins are the workhorses of the cell, performing a wide variety of functions. Most often, they perform these functions by interacting with other proteins, giving rise to large and complex protein-protein interactions networks.

In the last few years, the interest for the information retrieved from protein interactions increased and became one of the most important resources for

many scientific fields (drug discovery, disease treatment, biotechnology, medicine). Protein interaction maps are used to extract functional information and for the identification of metabolic or signal transduction pathways. These interactions are warranted by a wide range of protein binding interfaces ranging from long structured region (globular domains) (Richardson, 1981) to short unstructured region (short linear motifs (SLiMs)) (hunt, 1990). So far, globular domains have been studied more than SLiMs (more than 11000 domains are currently known) (Finn, et al., 2010), due to their long sequence length and also because their structure can be solved by x-ray crystallography.

### **1.1. Protein-Protein interaction modes**

Protein interactions are relatively balanced by their affinity for their various partners (proteins, ligands, nucleic acids, ions). Interaction specificity and strength are directly related to the protein sequence composition, the structure, the dynamics and the energetics (Neduva and Russell, 2006).

Proteins have several ways to interact with each other. In many cases the interactions are mediated by a protein component, which is commonly called a domain. Typically, a domain is a modular element with more than 36 residues (Jones, et al., 1998). In this case interactions between proteins often involve interaction between their domains (Yellaboina, et al., 2011), but in other cases the interaction is mediated by a domain of the first partner and a short linear motif of the second partner (Neduva and Russell, 2006). Domain-domain interactions between pairs of proteins can be studied via structural determination techniques or sequence similarity. In contrast, domain-linear motif interactions are more difficult to identify by sequence comparison.

Linear motifs are usually identified in experiments such as the yeast two-

hybrid experiment (von Mering, et al., 2002), but can also be detected computationally. The number of functional motifs so far discovered is much lower than the thousands of domains that could bind them. This suggests that there are probably hundreds of new motifs still to be discovered. In the next paragraphs I will outline all the main linear motif attributes and the methods and tools used to identify them.

## 1.2. Short Linear Motifs

### 1.2.1. What are Short Linear Motifs (SLiMs)?

**Short Linear motifs** (also known, as SLiMs, Linear Motifs (LMs) or minimotifs) are functional microdomains that mediate protein-protein interactions. Tim Hunt introduced the first definition for LMs in 1990 (hunt, 1990):

“The sequences of many proteins contain short, conserved motifs that are involved in recognition and targeting activities, often separate from other functional properties of the molecule in which they occur. These motifs are linear, in the sense that three-dimensional organization is not required to bring distant segments of the molecule together to make the recognizable unit. The conservation of these motifs varies: some are highly conserved while others, for example, allow substitutions that retain only a certain pattern of charge across the motif.”

The main features of linear motifs are:

- 1) They are **short** i.e generally composed by a short stretch of contiguous amino acids (3 to 10 residues), of which at least three are conserved, and mediate the binding. It is estimated that 70% of known linear motifs have 4 defined positions or less (Davey, et al., 2010). These are the positions that usually play an important role for the protein function (e.g., the consensus

motif PxxP binds to the SH3 protein domain (Pawson and Nash, 2003), where P is a Proline and x is any amino acid) (Figure 1). A position of a linear motif can be degenerated. This means that a functional residue can be substituted with another amino acid sharing similar physicochemical features without affecting motif functionality (Davey, et al., 2010).

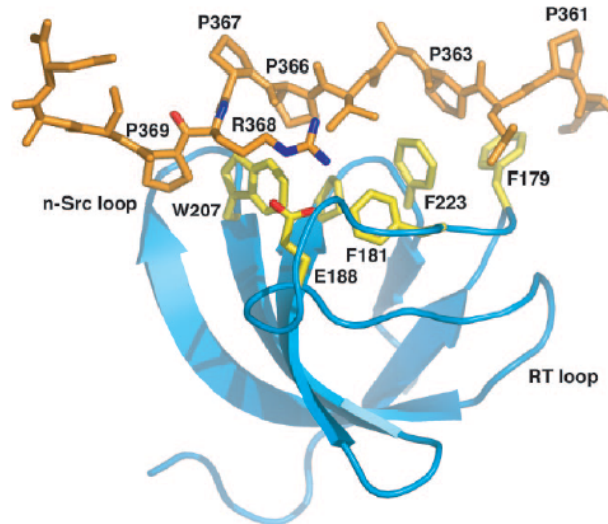
2) SLiMs are **linear** in the sense that the residues involved in the protein function are adjacent in the primary sequence and in close proximity in the tertiary structure.

In summary we can define **Motifs** as short sequence patterns, necessary for the protein function.

Linear motifs are observed in different proteins with different functions, such as transcription factors, adaptors, membrane receptors and mediators of protein-protein interactions (Linding, et al., 2007). We can classify SLiMs in four main functional classes (Diella, et al., 2008):

- Protein binding motifs: these are the most common linear motifs that bind to a domain of an interacting protein (PDZ binding motif, SH3 binding motif).
- Localization/targeting: represent a class of motifs involved in recognition of a protein to be targeted to a specific sub-cellular location (Nuclear export signal, ER retention retrieving).
- Cleavage: linear motifs acting as peptidase cleavage sites (sites of proteins that are cut by enzymes (Taspase 1, Furin).
- Post Translation Modification (PTM) sites: Some linear motifs are post-translational modification sites (N-linked glycosylation, N-Myristoylation, etc.).





**Figure 1:** The structure of SH3p40 in complex with the polyproline motif of p47;phox (PDB code 1W70). In blue is the core SH3p40 structure formed by two  $\beta$ -sheets. In orange is the p47phox polyproline peptide (residues 360–372). In yellow are the SH3p40 residues interacting with the polypeptide (Massenet, et al., 2005).

### 1.2.1.1. SLiMs biological features

In contrast to globular domains, linear motif functions are independent of their tertiary structure; most of them have been identified in disordered regions (~ 85% of occurrences), but they have been also observed in the accessible parts of globular domains (Russell and Gibson, 2008). The disorder tendency of LMs is important for the protein disorder-order transition upon motif-domain interactions. Residues in disordered regions are less evolutionary constrained than those in ordered regions and can more easily evolve towards functional motifs (Davey, et al., 2009). LMs are often enriched in some specific residues (e.g. R, K, P, W, T and C), and depleted of some others, especially A and G. Hydrophobic residues, I, M and V are interchangeable due to their chemical similarity in many SLiMs (Puntervoll, et al., 2003). Conserved motif residues are usually insufficient for high specific binding interactions; other residues surrounding the core motif are

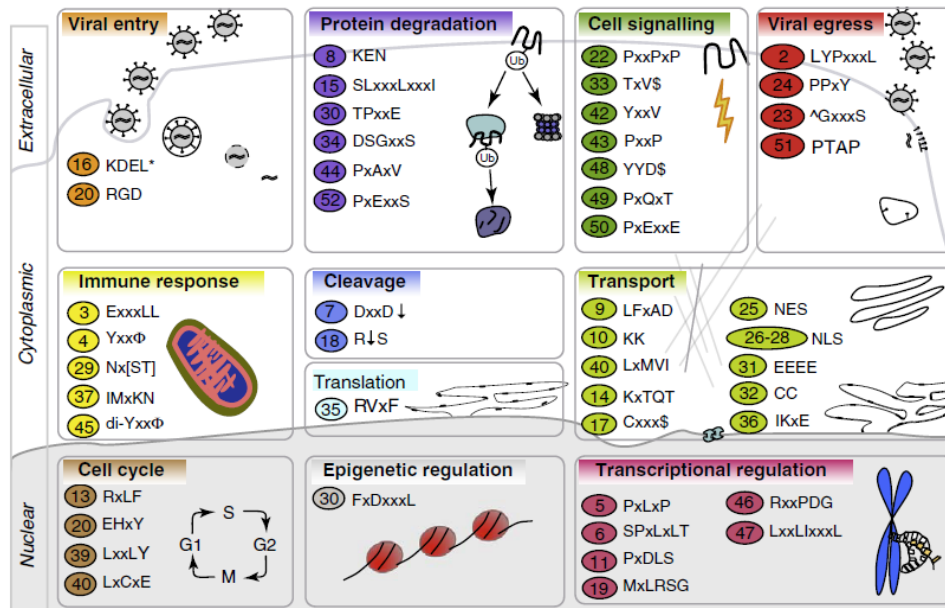
needed. SLiMs binding affinity is usually weak compared to those of domain-domain interactions, due to motif short length. The known motif features can be used to build new motif discovery tools and improve existing ones.

### **1.2.2. Why are we interested in Linear Motifs?**

We analyse LMs for several reasons. The most important one is that they are implicated in several disease pathways, essentially caused by motif mutations (such as Alzheimer's and Huntington's diseases (Davey, et al., 2011)). Since proteins use motifs for a wide range of functions, many viruses try to mimic human LMs to hijack cells. For example, the motif RxLx[QE] is implicated in the malaria pathogenesis and a disfunction of the SH3 binding motif is related to the Influenza, Hepatitis B and C viruses (Kadaveru, et al., 2008). A list of disease related LMs is reported in Figure 2.

The second reason is that linear motifs are very important for drug discovery. Recent studies have established the ability of small peptides to competitively bind proteins and the ability to target drugs to SLiM interactions (e.g. the angiogenesis inhibitor Cilengitide have provided promising results in cancer therapeutics (Burke, et al., 2002)).

The third reason is that, in spite of their importance, only a small number of them have been discovered. It is estimated that 15-40% of protein interactions are mediated by linear motifs. This implicates that hundreds of them are still to be discovered (Neduva and Russell, 2006). And this is the observation that inspired this thesis work.



**Figure 2:** Examples of diseases SLiMs, classified by function and cell compartment. For more details see “How viruses hijack cell regulation” (Davey, et al., 2011).

### 1.3. Linear motifs discovery

Linear motifs are difficult to detect experimentally or computationally due to their short length, low binding affinities, transient interactions, low evolutionary conservation, and because they can be easily activated or deactivated by a point mutation (Puntervoll, et al., 2003).

To date, linear motifs have usually been found by experimental methods, which are expensive and time consuming. Only in the last few years some computational tools were developed to analyse and predict them.

In an early attempt, many databases (e.g. the Eukaryotic Linear Motif resource (ELM) (Puntervoll, et al., 2003), and MiniMotifMiner (MnM) (Rajasekaran, et al., 2009)) tried to collect all linear motifs from the available literature and catalogue them according to their regular expression ([http://en.wikipedia.org/wiki/Regular\\_expression](http://en.wikipedia.org/wiki/Regular_expression)). A regular expression is a special text string made up of meta-characters, i.e. characters with a special

meaning capable of capturing motif information. An example is reported on page 22.

To enlarge the number of proteins expressing a functional linear motif, in a second attempt, many repositories used known linear motif patterns as a starting point to predict them.

Despite all the progress in this field, researchers still believe that the number of known SLiMs is too small. This implies that we need to develop new tools able to discover new linear motifs. In this regard, Short Linear Motif Finder (SLiMFinder) (Edwards, et al., 2007), and Discovery of Linear Motifs (DILIMOT) (Neduva and Russell, 2006) represent the most recent and accurate *de novo* SLiMs discovery tools.

Computational methods to linear motif discovery can be grouped in two categories:

- 1) Methods aimed at identifying new instances of already known short linear motifs in proteins.
- 2) Methods aimed at discovering new / *de novo* linear motifs.

In Chapters 2 and 3 I will outline the main concepts, data, and tools used to analyze and discover linear motifs for each of these two categories.

#### **1.4. Aim and contributions of the study**

The purpose of the study presented here is to contribute to the field of bioinformatics by developing, testing and applying computational methods to discover new short linear motifs. Our original results are described in three independent papers, which are briefly described below.

Paper I “ELM: the status of the 2010 eukaryotic linear motif resource” (Gould, et al., 2010) describes a knowledge base covering 159 known linear motifs, and more than 1300 experimentally reported instances. This resource

is implemented in a web-server accessible at the URL (<http://elm.eu.org/>). It also includes a tool for predicting new candidates of known linear motifs in user-defined protein sequences (Gould, et al., 2010).

Paper II “Exploiting Publicly Available Biological and Biochemical Information for the Discovery of Novel Short Linear Motifs” (Sayadi, et al., 2011). In this paper we describe a novel approach for the discovery of SLiMs based on their occurrence in evolutionarily unrelated proteins belonging to the same biological, signaling or metabolic pathway and give specific examples of its effectiveness in both rediscovering known motifs and in discovering novel ones. An automatic implementation of the procedure, available for download, allows significant motifs to be identified, automatically annotated with functional, evolutionary and structural information and organized in a database that can be inspected and queried. An instance of the database populated with pre-computed data on seven organisms is accessible through a publicly available server and we believe it constitutes by itself a useful resource for the life sciences (<http://www.biocomputing.it/modipath>) (Sayadi, et al., 2011).

In paper III “The interaction network of the 14-3-3 protein in the ancient protozoan parasite *Giardia duodenalis*” (Paper revision submitted to *Journal of Proteome Research*, Lalle, et al., 2011) we describe a combined experimental and computational study of the interaction network of the protein g14-3-3 in the ancient protozoan parasite *Giardia duodenalis*. This work was performed in collaboration with the group of Dr. Marco Lalle at *Istituto Superiore di Sanità* (ISS).

## 2. Methods aimed at identifying new instances of known LMs

Various projects have tried to collect information about linear motifs through experimental methods, extensive literature searches and high-throughput data analyses. Many repositories storing known LMs are available such as PROSITE (Bairoch, 1993), ELM (Puntervoll, et al., 2003) and MnM (Rajasekaran, et al., 2009). Some of them are focused on collecting specific type of linear motifs, such as phosphorylation sites (e.g. Phospho.ELM (Diella, et al., 2004)) or cleavage sites (MEROPS (Rawlings, et al., 2008) and CutDb (Igarashi, et al., 2007)). All these resources have contributed to highlight many LMs features, such as their evolutionary conservation or their propensity to occur in disordered regions. These motif attributes were used by the developers of several tools to identify new instances of known linear motifs.

The QuasiMotifFinder tool (Gutman, et al., 2005) searches for conserved motifs in proteins using PROSITE patterns. It assigns to motif occurrences a score based on their physico-chemical similarity to the original motif and on the degree of evolutionary conservation of the residues appearing in the occurrence within homologous sequences. This method is restricted to the set of motifs in PROSITE.

The AutoMotif Server (AMS) (Plewczynski, et al., 2005) predicts PTM sites in proteins, based only on sequence information. Unfortunately, the biological significance of the score assigned to the predicted instances is not clear.

Such tools helped considerably to detect many new motif instances, but they still suffer from over-prediction problems and limitations in their application. To overcome these problems, in a recent work (Gould, et al., 2010) we gave some hints aimed at helping collecting new SLiMs and detecting novel instances of known functional motifs.

### **2.1. ELM: The Eukaryote linear motif resource**

As previously mentioned (Gould, et al., 2010), the ELM resource represents an expanded knowledge base (<http://elm.eu.org>) that stores manually curated information about known linear motifs (159 motifs and lists more than 1300 instances; Instances are proteins bearing a linear motif). It also provides a tool to search for matches of known motifs in protein sequences. This resource was created to contribute to LM discovery and to help researchers select good candidates for experimental validations. The ELM resource has become available online in 2003 (Puntervoll, et al., 2003).

The main problem of new instance detection resides essentially in the motif short length, which leads to a high number of false positive matches. This happens because a string of few characters has a high probability of being found by chance in a long sequence of characters. As a solution to this, in the last few years it has become clear that selective criteria should be applied to reduce the number of false positives and as a consequence, make the detection and the analysis of new instances of known linear motifs more meaningful from a biological point of view.

This can be done by using LM features such as evolutionary conservation, localization in disordered regions and accessibility to the solvent. Cellular localization is also important for the functional likelihood of the motif (Davey, et al., 2011). Using these features, we have implemented several logical filters in the ELM server: Conservation score filter (CS) (Chica, et

al., 2008), Structure filter (Via, et al., 2009), taxonomic filter, cell compartment filter and SMART globular domain filter (Letunic, et al., 2009).

### 2.1.1. Results

ELM linear motifs are usually manually collected using extensive literature searches. Recently a new tool called MiMosa (Vyas, et al., 2010) was created to facilitate the literature tracking and annotation of new linear motifs.

Few initial steps are necessary before adding a new entry to the ELM database. We start by mining literature papers in search for evidences of experimentally determined motifs. When they are found, we review them for quality issues, type of experiment and perform a number of computational tests to establish whether the motifs meet our criteria for true positives, i.e. they are evolutionary conserved (Chica, et al., 2008), exposed to the solvent (Dosztanyi, et al., 2005) and localized in the suitable cell compartment (Gould, et al., 2010). Instances that pass all the tests are used to derive a regular expression that best describes the motif. Finally, the collected motif information is stored as a new entry of the database.

Once a new motif is added to the ELM database we try to detect its new instances. This is a difficult task, being the signal too weak to know whether an occurrence in a protein is true or it just appears by chance. We found that the search is more successful if we limit the search space as much as possible. Accordingly, interaction data and motif enrichment in proteins with annotated keywords were implemented to provide some statistical support. Such approach has proved its efficiency in various cases, for example new instances of EH1 transcriptional repressor motif were enriched in proteins annotated with keywords related to transcription (Copley, 2005). In a recent



case, a tool called SIRW (Ramu, 2003) allowing motif/keyword explorations was used to detect new ELM instances of the KEN box APCC-binding Destruction motifs found to be enriched with cell cycle keywords (Michael, et al., 2008).

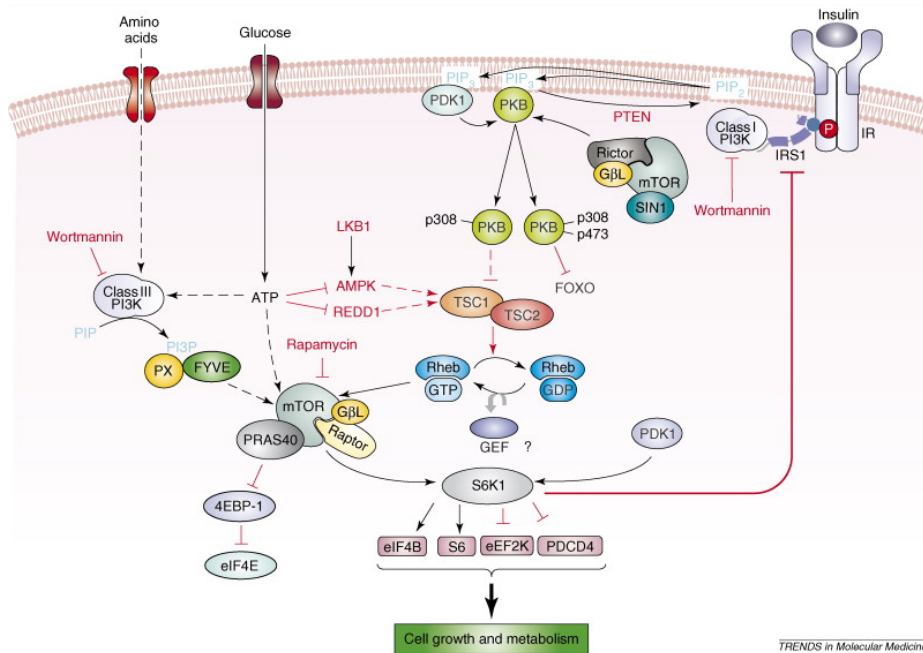
#### **2.1.1.1. mTOR related motifs**

In this paragraph, we want to elucidate the ELM procedure, by describing the example of the mTOR target motifs. These are linear motif known to mediate some of the mTOR protein interactions.

The mTOR protein, the mammalian target of rapamycin (Schalm, et al., 2003), is a serine/threonine kinase that participates to the mTOR pathway, which plays a key role in the regulation of cell growth, cell proliferation, protein synthesis, and transcription (Hay and Sonenberg, 2004). It is dysregulated in many human diseases, especially in certain cancers, in diabetes and obesity (Beavers, et al., 2006). There are two functionally distinct mTOR-containing multiprotein complexes. In the first complex, TORC1 which is rapamycin-sensitive, mTOR acts by phosphorylating and inactivating the eukaryotic initiation factor 4E-binding protein (4E-BP1) and by phosphorylating and activating the S6 kinase (S6K1) (Schalm, et al., 2003). In the second complex, TORC2 which is rapamycin-insensitive, mTOR acts by phosphorylating the protein kinase B (PKB) (Chiang and Abraham, 2007).

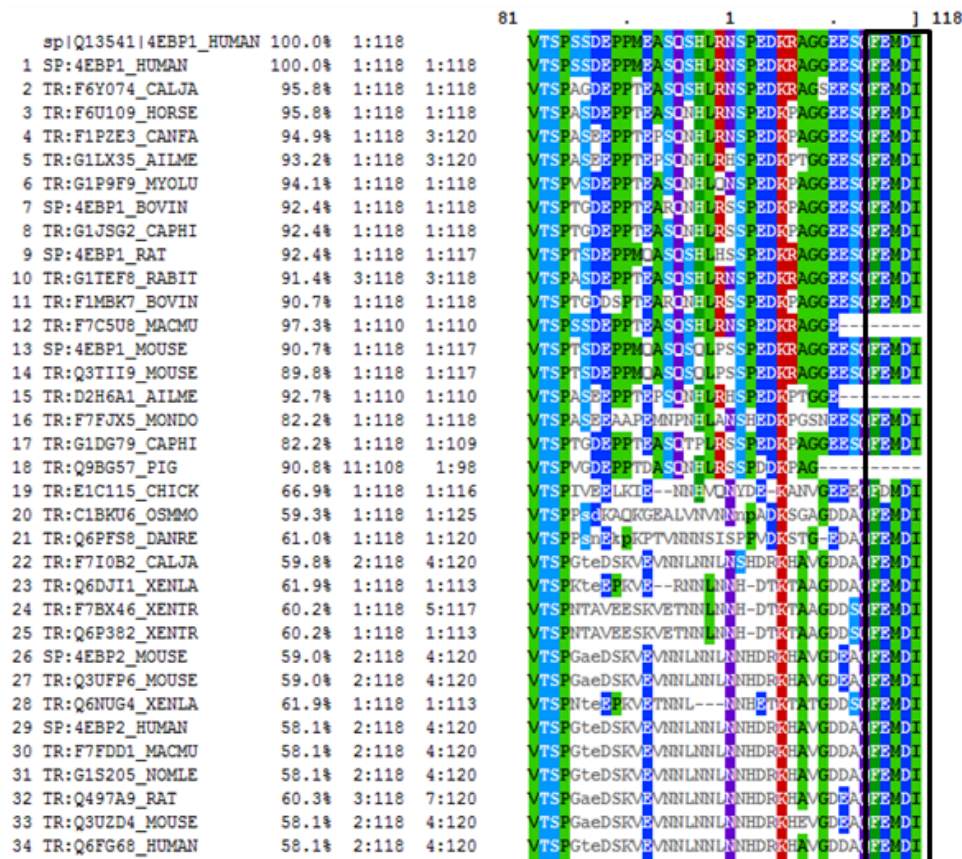
Two short motifs have been reported in the mTOR target proteins: the TOS motif (TOR signaling motif) and the RAIP motif. These mediate interactions respectively in the TORC1 and TORC2 complexes. mTOR in complex with an adaptor protein (raptor) phosphorylates its substrates S6K1 and 4E-BP1 in the TORC1 complex through the TOS motif, present in the N-terminus of S6K1 and C-terminus of 4E-BP1. In contrast, in the TORC2 complex,

mTOR uses the RAIP motif to phosphorylate the N-terminus of 4E-BP1 (Beugnet, et al., 2003). A detailed figure (Figure 3) of the mTOR pathway is showed below (Chiang and Abraham, 2007).



**Figure 3:** The mTOR pathway (Chiang and Abraham, 2007).

To analyse these two motifs, we collected several experimental papers describing TOS and RAIP instances (Beugnet, et al., 2003; Carroll, et al., 2006). We found eight putative instances of the TOS motif (Beugnet, et al., 2003). For each instance, we verified the biological significance, sequence disorder, evolutionary conservation (Figure 4), and quality of the experiment. In our procedure, motif existence is taken under further consideration only if it has been validated by at least three different experimental methods. For example, four methods were used to gain evidence on the motif “FEMDI” in the protein 4EBP1 (Beugnet, et al., 2003): co-immunoprecipitation, western blot, pull down and mutation analyses (Berggard, et al., 2007).



**Figure 4:** Multiple sequence alignment of the 4E-BP1 homologous proteins obtained with MView (Brown, et al., 1998). The box in black shows the conservation of the motif FEMDI along the homologous proteins.

Only five out of the eight putative TOS instances passed our tests and were therefore considered as true positives. In contrast the two proteins interacting with mTOR through the RAIP motif were considered as false positives (Carroll, et al., 2006). All potential mTOR instances are listed in Table 1.

Motif	Uniprot ID	Sequence	Position	
TOS	4EBP1_HUMAN	FEMDI	114-118	C-Ter
	4EBP2_HUMAN	FEMDI	116-120	C-Ter
	4EBP3_HUMAN	FEMDI	96-100	C-Ter
	KS6B1_HUMAN	FDIDL	28-32	N-Ter
	KS6B2_HUMAN	FDIDL	5-9	N-Ter
	HIF1A_HUMAN	FVMVL	99-103	
	PLD2_HUMAN	FVQLF	257-261	
	AKTS1_HUMAN	FVMDE	129-133	
RAIP	4EBP1_HUMAN	RAIP	13-16	N-Ter
	4EBP2_HUMAN	RAIP	15-18	N-Ter

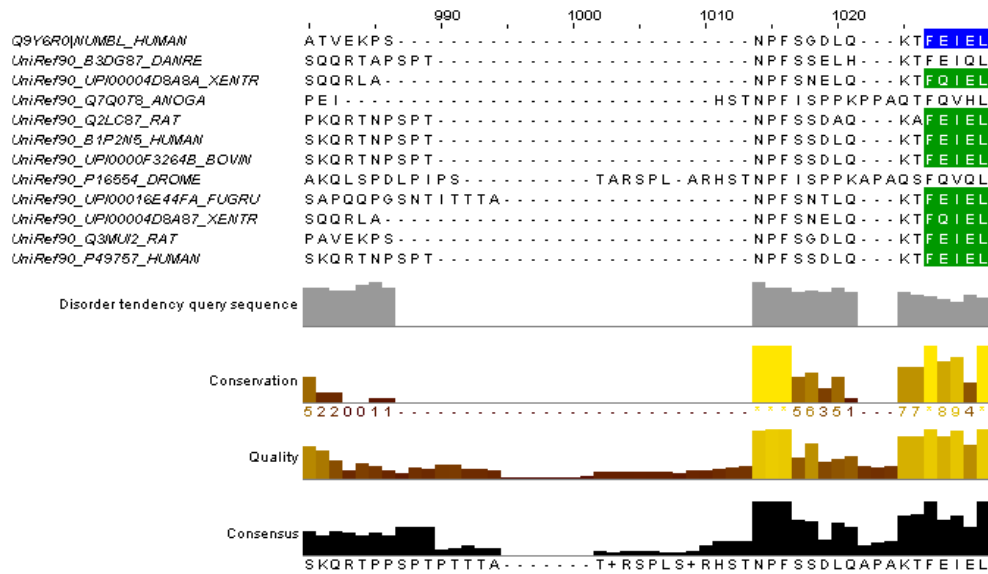
**Table 1:** TOS and RAIP putative instances.

The multiple sequence alignments of TOS instances were used to identify functional important residues and therefore to derive their representative regular expression:

**F[EDQS][MILV][ED][MILV]({0,1}[ED])(\$)**

Where ‘.’ means any amino acid, ‘[EDQS]’ means that all amino acids between [ ] are allowed, it can be ‘E’ or ‘D’ or ‘Q’ or ‘S’. ‘|’ means ‘or’, i.e. X|Y means that either X or Y can match. Curly brackets indicate a range of allowed positions, i.e. ‘{0,1}’ means 0 is the min required, and 1 is the max allowed. ‘\$’ means that the motif matches the carboxy-terminus of a protein sequence. From this regular expression we can deduce that the motif alternates between hydrophobic and polar positions and its terminal charge can be either in the carboxy-terminus (4E-BP1,2,3) or inside the protein sequence (S6-beta kinase). The motif regular expression and the related information give rise to a new entry in the ELM database. The TOS motif is stored under the ELM ID: LIG\_RAPTOR\_TOS\_1 accessible at [http://elm.eu.org/elms/elmPages/LIG\\_RAPTOR\\_TOS\\_1.html](http://elm.eu.org/elms/elmPages/LIG_RAPTOR_TOS_1.html) (Figure 6). The newly added TOS motif was then used to search for new instances of the TOS motif using the Motif/keyword search method (Michael, et al., 2008). We identified two motif occurrences in NUMB\_HUMAN and

NUMBL\_HUMAN, respectively (Figure 5), and evaluated them with the ELM logical filters (e.g. Conservation Score (CS) (Chica, et al., 2008)) (Figure 7).



**Figure 5:** CS web interface result output, displayed with the annotated sequence alignment editor JalView (Waterhouse, et al., 2009). The alignment shows the set of sequences obtained by the CS filter with the NUMBL query sequence at top position. Sequences that align to the TOS motif regular expression are colored in green showing that the motif is well conserved, with a top CS score of 1.00. CS gives a score varying from 0 to 1, where 0 means not conserved and 1 well conserved (Chica, et al., 2008).

**ELM** The Eukaryotic Linear Motif resource for *Functional Sites in Proteins*

Search ELMs Instances Candidates Links About News Help Diseases Viruses

**LIG\_RAPTOR\_TOS\_1** 246

<< LIG\_PTB\_Phospho\_1 << Menu >> LIG\_Rb\_LxCxE\_1 >>

**Functional Site Class:** Raptor interacting motif

**Functional site description:** At least one (but perhaps more) motifs interact with raptor as part of TOR pathway function. TOR is a kinase that regulates translation in a pathway with broad effects on many other cellular processes. The best known raptor-binding motif is TOS (the TOR signalling motif).

**ELMs:** LIG\_RAPTOR\_TOS\_1

**Description:** The TOS motif is found in substrates of the mTOR kinase. TOS interacts with the WD40-containing adaptor protein Raptor that is required to bring TOR together with its substrates. This motif expression is derived from the conservation observed in the 4E-BP translation initiation factors and the ribosomal S6-beta kinases and shows a strong beta amphipathicity. The motif alternates between hydrophobic and negatively charged residues. Although the structure of a TOS-Raptor complex is not yet known, the conservation might imply that the motif is bound in a sandwiched pocket. The terminal charged moiety can be either the carboxy-terminus (4E-BP1,2,3) or an amino acid (S6-beta kinase). TOS motifs reported in Hif1-alpha, Plg2 and AKTS1 do not match the motif expression in the ELM resource. The TOS motif is clearly not present in many proteins but the NUMB and NUMBL proteins do contain plausible candidates that might be worth investigating. The TOS motif as described here is found in metazoa, slime mould and one basidiomycete, but not in other fungi. If the phylogenetic distribution is larger, then the TOS motif will have diverged in different lineages.

**Pattern:** F[EDQS][MLV][ED][MLV]((.{0,1}[ED])|(\$)) (Probability: 0.000207)

**Present in taxons:** Eukaryota

**Interaction Domain:** WD40 (PF00400)  
WD domain, G-beta repeat  
(Stoichiometry: 1:1)

■ See 5 Instances for LIG\_RAPTOR\_TOS\_1

■ **Abstract**

mTOR the mammalian target of rapamycin, also known as FRAP, is a member of a superfamily of protein serine/threonine kinases termed PIK-related (Oshiro, 2007), that was first identified in *Saccharomyces cerevisiae*. The TOR pathway is an emerging target for the treatment of cancer, diabetes and obesity (Eguchi S, 2007).

TOR controls protein synthesis through a stunning number of downstream targets. Some of the targets are phosphorylated directly by TOR, but many of them are phosphorylated indirectly. TOR participates in at least two distinct multiprotein complexes TORC1 and TORC2. These complexes play important role in the regulation of cell growth, cell proliferation, cell motility, cell survival, protein synthesis, and transcription (Beugnot A, 2003).

Two short motifs have been reported in the targets of TOR, the TOS motif (TOR signalling motif) and the RAIP motif. (The RAIP motif is not characterised well enough to be included in ELM).

In mammals mTOR, (mammalian TOR), regulates protein synthesis through the phosphorylation and inactivation of the repressor of mRNA translation, eukaryotic initiation factor 4E-binding protein (4E-BP1), and through the phosphorylation and activation of S6 kinase (S6K1). mTOR exists in two distinct complexes within cells: one (mTORC1) that contains mTOR, mLst8 and raptor and another (mTORC2) containing mTOR, mLst8, mSin1 and rictor (Schalm S, 2002).

In order to phosphorylate its substrates in the TORC1 complex, mTOR needs an adaptor protein (raptor), that binds to S6K1 and 4E-BP1 (Eguchi et al., 2006). The interaction of raptor with S6K1 and 4E-BP1 is mediated by the TOS motif that is present in the N-terminus of S6-beta kinases and C-terminus of 4E-BP1.

The TOS motif is currently only known from metazoan organisms. However, the TORC1 complex and highly conserved raptor orthologues are present in both budding yeast (*Kog1*) and fission yeast (*Mip1*). It is possible that the TOS motif exists in other Eukaryotic lineages but diverges from the metazoan motif pattern and so needs rediscovery.

■ 10 selected references: Show

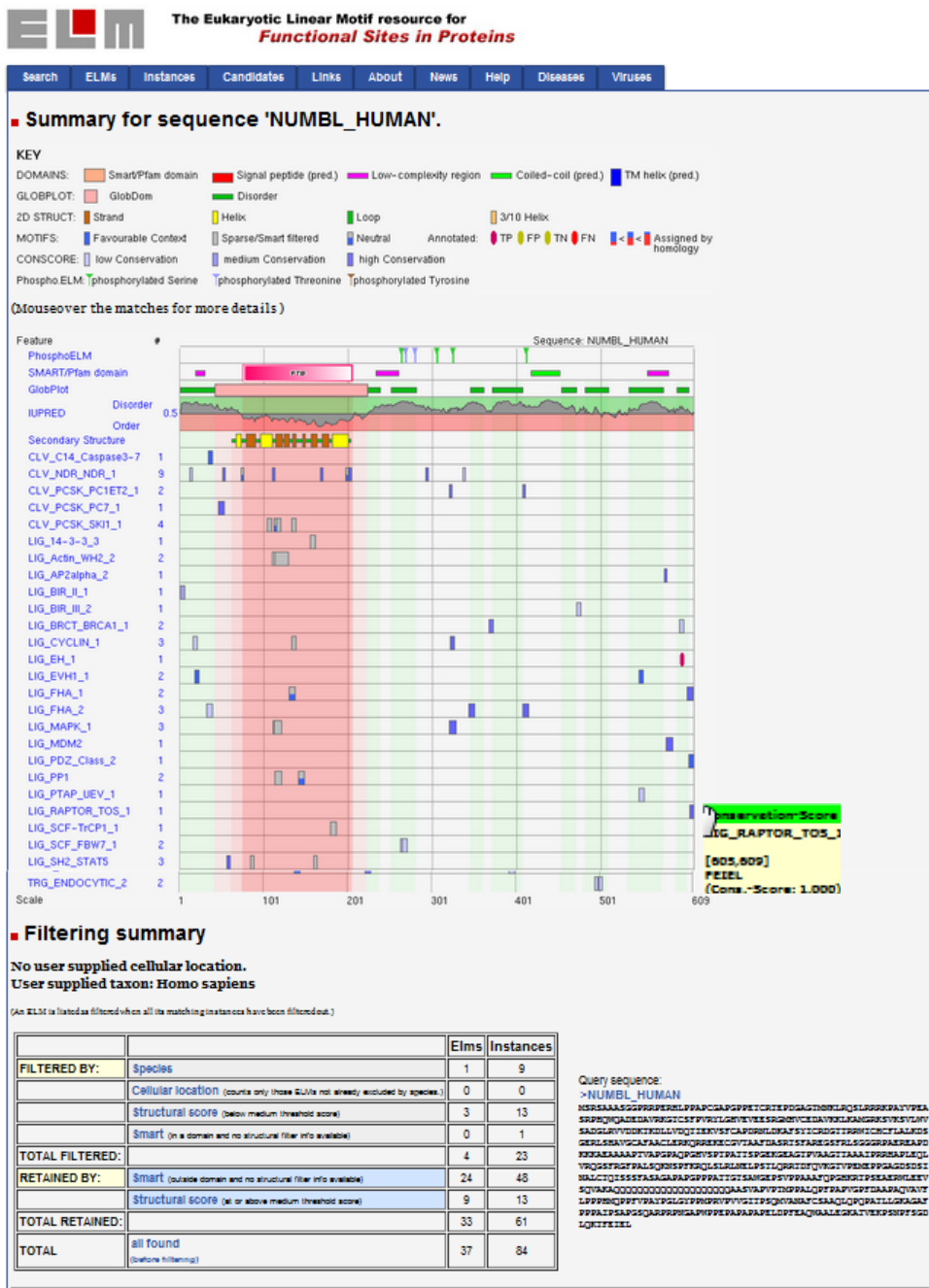
■ 4 GO-Terms: Show

■ 5 Instances for LIG\_RAPTOR\_TOS\_1  
(click table headers for sorting)

Sequence	Start	End	Subsequence	Instance Logic	PDB	Organism
4EBP1_HUMAN	114	118	RNSPEDKRAGGEESQ <b>FEMDI</b>	true positive	---	<i>Homo sapiens</i> (Human)
4EBP3_HUMAN	96	100	LKEQETEEEIPDDAQ <b>FEMDI</b>	true positive	---	<i>Homo sapiens</i> (Human)
4EBP2_HUMAN	116	120	LNNHDRKHAVGDDAQ <b>FEMDI</b>	true positive	---	<i>Homo sapiens</i> (Human)
KS6B1_HUMAN	28	33	AEDMAGV <b>FDLDL</b> DQPEDAGS	true positive	---	<i>Homo sapiens</i> (Human)
KS6B2_HUMAN	5	10	MAAV <b>FOLDLE</b> TEEGSEGECE	true positive	---	<i>Homo sapiens</i> (Human)

**Figure 6:** ELM web page details of the entry LIG\_RAPTOR-TOS\_1 ([http://elm.eu.org/elms/elmPages/LIG\\_RAPTOR\\_TOS\\_1.html](http://elm.eu.org/elms/elmPages/LIG_RAPTOR_TOS_1.html)). Information about the function, the description and the Pattern of the motif are shown first, the list of proteins, motif occurrences and positions are shown at the end of the web page (Gould, et al., 2010).





**Figure 7:** Output page of the ELM server when queried with the protein sequence NUMBL (UniProt: NUMBL\_HUMAN). Each key refers to specific information, e.g. green indicates domains. Structural, sequence disorder and other information are also provided (Gould, et al., 2010).

### 3. *de novo* Linear motif prediction

During the last few years, many proteome-scale interaction data sets have become available, opening the door to the development of new linear motif discovery algorithms (e.g. DILIMOT (Neduva and Russell, 2006), SLiMdisc (Davey, et al., 2006) and MOVIN (Marcatili, et al., 2008)).

*De novo* SLiM discovery tools use the concept of statistical over-prediction as an indicator of functionality. Any set of proteins that are likely to use SLiMs to mediate their functionality represents a suitable input for such tools. The majority of *de novo* SLiM discovery tools is based on protein-protein interaction data, usually obtained in high-throughput experiments and from Protein-protein interaction (PPI) databases (e.g. STRING (Jensen, et al., 2009), MINT (Zanzoni, et al., 2002); BIND (Alfarano, et al., 2005)).

Proteins in a network could potentially use SLiMs to interact. If this is the case, SLiMs might be searched in sub-networks, i.e. in sets of proteins ( $\geq 2$ ) interacting with a single central protein. Statistically, the signal from proteins that interact through a motif must be stronger than the noise due to proteins that either use a different interactions mechanism or are false positives. This means that PPI data quality is essential in motif discovery from PPI networks.

In a recent work we proposed a new approach where, instead of using PPI data to discover motif, we used sets of proteins belonging to the same biological, signaling or metabolic pathway. This approach is described in paper II “Exploiting Publicly Available Biological and Biochemical Information for the Discovery of Novel Short Linear Motifs” (Sayadi, et al., 2011) (enclosed in this thesis).



### 3.1. MoDiPath: Motif Discovery in Pathways

A linear motif can be derived from a set of proteins sharing a function or interacting with a common central protein. Therefore, by collecting and analyzing the sequences of proteins sharing common features, it should be possible to identify novel motifs associated with the common functionality. In this work we aimed at discovering motifs associated with biological processes. To this end, we inspected sequences belonging to the same process (Yaffe, et al., 2001). As described in paper II, we combined different databases and tools to generate a new computational approach called MoDiPath (Motif Discovery in Pathways (MDP)) that uncovers new linear motifs over represented in pathways. The database is accessible online at <http://www.biocomputing.it/modipath>.

MoDiPath is designed to look for over-represented motifs in pathways. Our approach consists in taking a set of proteins belonging to a pathway (metabolic or signaling transduction), removing sequence redundancy, and applying algorithms for the identification of motifs significantly over-represented in the whole set or in a subset. Statistical tests and other filters are then applied to check the robustness of the predicted motifs (Sayadi, et al., 2011).

#### 3.1.1. Results

Our source of pathway information is the KEGG (Kyoto Encyclopedia of Genes and Genomes) Pathway database, which is a knowledge base for systematic analysis of gene functions (Ogata, et al., 1999). It clusters proteins in pathways for several species. Each pathway represents the functional aspects of a biological system, and involves a specific protein list, graphically represented as a network of connected proteins (Kanehisa and

Goto, 2000).

In this work, the KEGG database was used to derive information from seven different species (*Homo sapiens*, *Mouse*, *Rat*, *Drosophila*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, *Escherichia coli*). As an example, here we report the results obtained for *Homo sapiens*.

The KEGG database counts more than 201 pathways for the human species. Each pathway contains a variable number of proteins. In our procedure each pathway protein set is screened in search for shared motifs using SlimFinder, a *de novo* linear motif discovery tool (Edwards, et al., 2007). To avoid that high intra-pathway sequence similarity affects motif prediction, we first remove redundancy at 25% and 40% sequence identity from each protein set using CD-HIT (Li and Godzik, 2006).

As a result, we obtain a list of predicted motifs for each pathway in the form of a regular expression. Our approach also provided a list of proteins for each predicted SLiM. A total of 2097 putative SLiMs were predicted in the *Homo sapiens* pathways. Predicted SLiMs were ranked according to their statistical significance. A hyper-geometric distribution test (Romero, et al., 2001) was used to assess whether the motif is statistically significantly over-represented in a pathway. To choose the best threshold for the hyper-geometric distribution p-value, we randomly reshuffled the protein list of each pathway ten times. Comparing the results of the real data with the random ones, we selected a p-value of  $3 \times 10^{-9}$ , as the best threshold, which corresponds to a false discovery rate (FDR) lower than 10%. According to this p-value, 104 out of the 2097 motifs were statistically significant (21 in metabolic pathways and 83 in non-metabolic pathways). This is a high number if we consider that the ELM database stores 159 LMs. We also verified whether the newly discovered motifs showed some degree of

similarity to known motifs with known SLiMs. We found that out of the 104 significant motifs, 82 have some degree of similarity to already known motifs present in other databases (e.g. ELM, MnM). Sixty-three of the eighty-two are identical to other known motifs. Interestingly, 22 SLiMs do not share any similarity with any known motif and can be considered as novel motifs. We define two motifs to be similar if their CompariMotif (Edwards, et al., 2008) score is above 0.7. CompariMotif is a tool that compares motifs between each other and gives a score of similarity ranging from 0 (weak similarity) to 1 (Strong similarity) (see Materials and Methods).

These results reflect the ability of the MoDiPath procedure to uncover and re-discover a significant number of motifs. A summary of the results for the seven organisms is reported in Table 2. Finding over-represented motifs in a pathway is crucial but not sufficient. It is important to distinguish between true and false occurrences, i.e. to assess which occurrences have a biological role and are associated to a biological function. To this aim, we used additional tools to carry out further investigations. We analyzed, for each motif, its degree of sequence disorder, its evolutionary conservation score (CS), structural information and GO term enrichment.

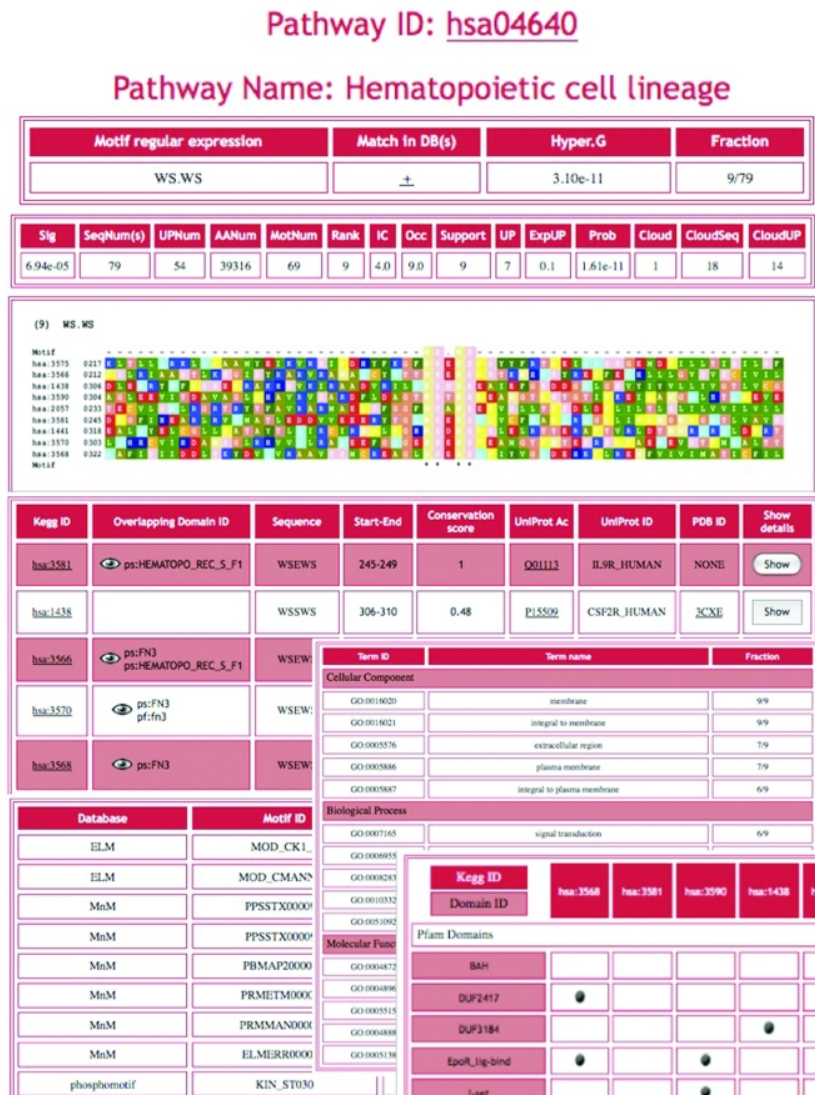
Species	Putative <sup>(a)</sup>			Significant SLiMs <sup>(b)</sup>			Novel SLiMs <sup>(c)</sup>		
	Total	MP	NMP	Total	MP	NMP	Tot	MP	NMP
<i>H.sapiens</i>	2097	836	1261	104	21	83	22	6	16
<i>M.musculus</i>	2094	882	1212	127	38	89	28	12	16
<i>R.norvegicus</i>	1863	809	1054	72	19	53	15	5	10
<i>D.melanogaster</i>	1391	632	759	35	5	30	4	0	4
<i>C.elegans</i>	1050	610	440	32	12	20	6	6	0
<i>E.coli</i>	933	733	200	11	10	1	2	1	1
<i>S.cerevisiae</i>	889	584	305	20	15	5	3	2	1

**Table 2:** (a): Total number of putative motifs identified by SliMFinder in KEGG pathways; (b): number of significantly over-represented motifs in pathways with respect to the two reference datasets (hyper-geometric p-value  $< 3 \times 10e^{-9}$ ); (c): number of significant motifs that are novel (hyper-geometric p-value  $< 3 \times 10e^{-9}$ , NormIC  $< 0.7$ ). MP: Metabolic pathways; NMP: Non-Metabolic Pathways.

As described in paper II, all these steps are part of an automatic pipeline that can be downloaded at <http://www.biocomputing.it/modipath/MoDiPath.11-04-2011.zip> and installed and run locally.

Results are stored in the MoDiPath web server (<http://www.biocomputing.it/modipath/>) and can be displayed via a visual interface (Figure 8). In the ‘Search’ part the user can access the available data selecting a protein ID, a pathway ID or a species name from the available species menu. In the ‘Scan’ part there are two options: Protein/Sequence Scan and Pattern Expression Scan. In the ‘Protein/Sequence Scan’ part, users are required to select a species of interest and to input a protein ID or a sequence in Fasta format. The program will scan the protein with all the motifs collected for the selected species. As a result, users obtain a list of motifs and their sequence position in the protein also mapped, when possible on their three-dimensional structure. In the ‘Pattern Expression Scan’ part, users are required to input a motif and to select a species; the program will provide a list of similar motifs present in the database, ranked by a similarity score. Computationally, the jobs run in a few seconds.

As an example of the application of MoDiPath, we now describe a case of a rediscovered motif (SKL\$) and of a newly discovered motif ([FL].L.C..Y..A).



**Figure 8:** The information provided by MoDiPath for the hsa04640 KEGG pathway. (a) First column: the SLiM regular expression; Second column: a ‘+’ is reported if the motif overlaps to a similar motif in other databases (the list of which is shown by moving the mouse over the ‘+’); Third column: the hyper-geometric p-value of the number of motif hits in the hsa04640 pathway compared to the number of motif hits in the SwissProt database; Fourth column: The fraction of proteins in the hsa04640 pathway that contain the WS.WS motif (b) Multiple sequence alignment of the hsa04640 pathway proteins containing the WS.WS motif. (c) Information about each of the hsa04640 proteins containing the WS.WS motif. Clicking on the ‘Show’ button provides more detailed information, including the protein structure visualization with the motif hit(s) highlighted. (d) List of motif overlap(s) to similar motifs in other databases; the last column reports the CompariMotif similarity score (NormIC). (e) GO terms shared by the hsa04640 pathway proteins that have the motif; the last column reports the fraction of the proteins hosting the motif that share a GO term.

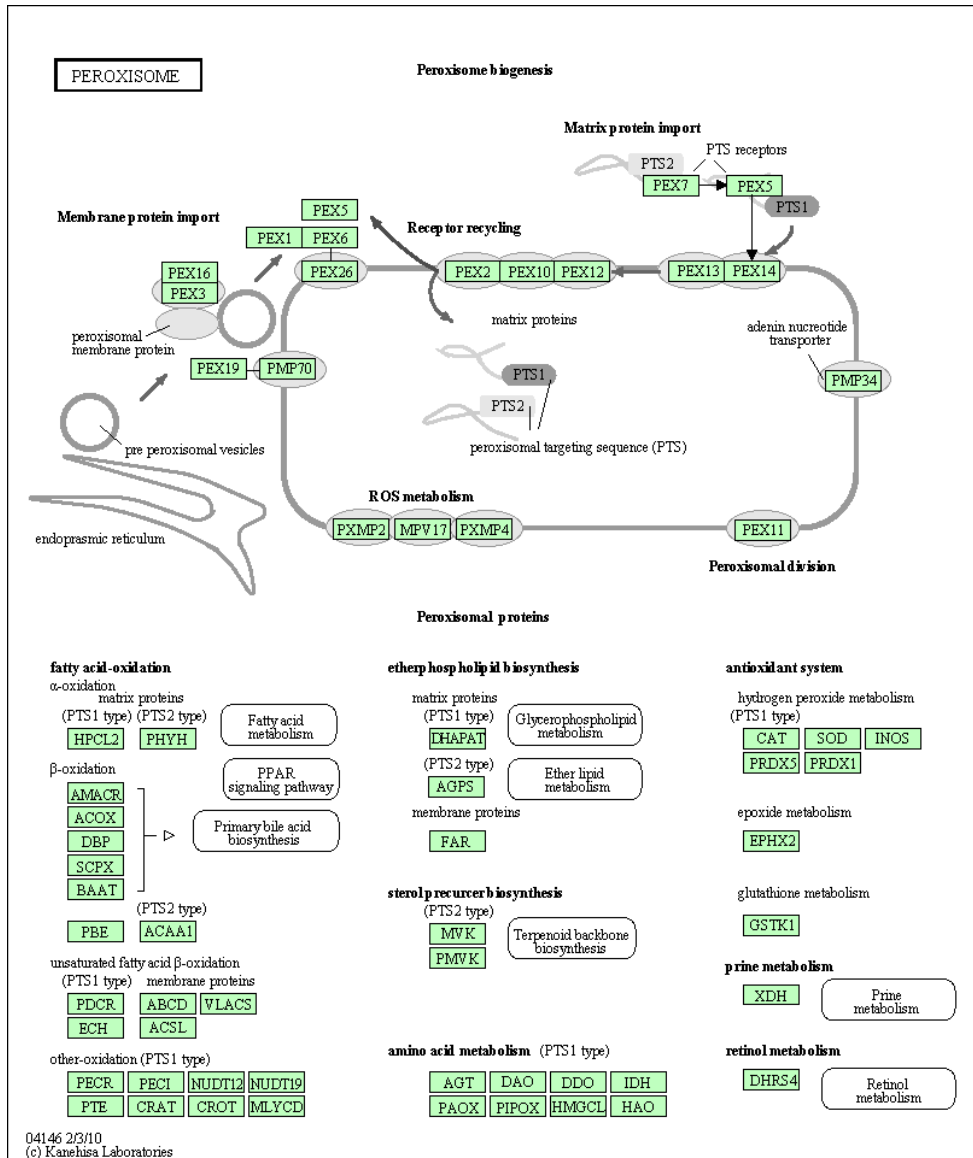
### 3.1.1.1. Rediscovered motif: “SKL\$”

MoDiPath allowed the rediscovery of the well-established motif “SKL\$”, which we found to be significantly over-represented (hypergeometric p-value,  $1.7 \times 10^{-11}$ ) in the Human Peroxisome pathway (KEGG ID: hsa04146). SKL (Serine-Lysine-Leucine) is a conserved tripeptide, found at the carboxy-terminal of protein sequences. The “SKL\$” motif is identical to the MnM motif annotated as Pex5-binding and associated to trafficking to Peroxisomes. The same annotation is reported for a similar motif in the ELM database (TRG\_PTS1, regular expression:  $(.[SAPTC][KRH][LMFI]S)$ ), which is annotated as a C-terminal signal interacting with the Pex5p protein to target proteins to the peroxisomal matrix.

Peroxisomes, are vascular organelles bounded by a single membrane and found in eukaryotic cell (Gabaldon, 2010). They play a key role in many metabolic processes such as fatty acid oxidation, metabolism of cholesterol and biosynthesis of ether-glycolipids (Gabaldon, 2010). Peroxisomal proteins are first synthesized in the cytosol and then imported into the Peroxisomes (Wendland and Subramani, 1993). The transport of proteins into the matrix of Peroxisomes is mediated by the motif “SKL\$”, known as peroxisomal targeting signal (PTS1). Peroxisomal cargo proteins, called peroxins (e.g. PEX5), recognize, tag and import proteins owing a PST1 motif into the Peroxisome (Saleem, et al., 2006) (Figure 9). A loss of this motif has been shown to imply a loss of peroxisomal functions and the consequent appearance of a defect known as peroxisomal disorder (PD). Patients with PD display severe neurological, hepatic, and renal abnormalities (Wendland and Subramani, 1993). This highlights the importance of studying the “SKL\$” motif.

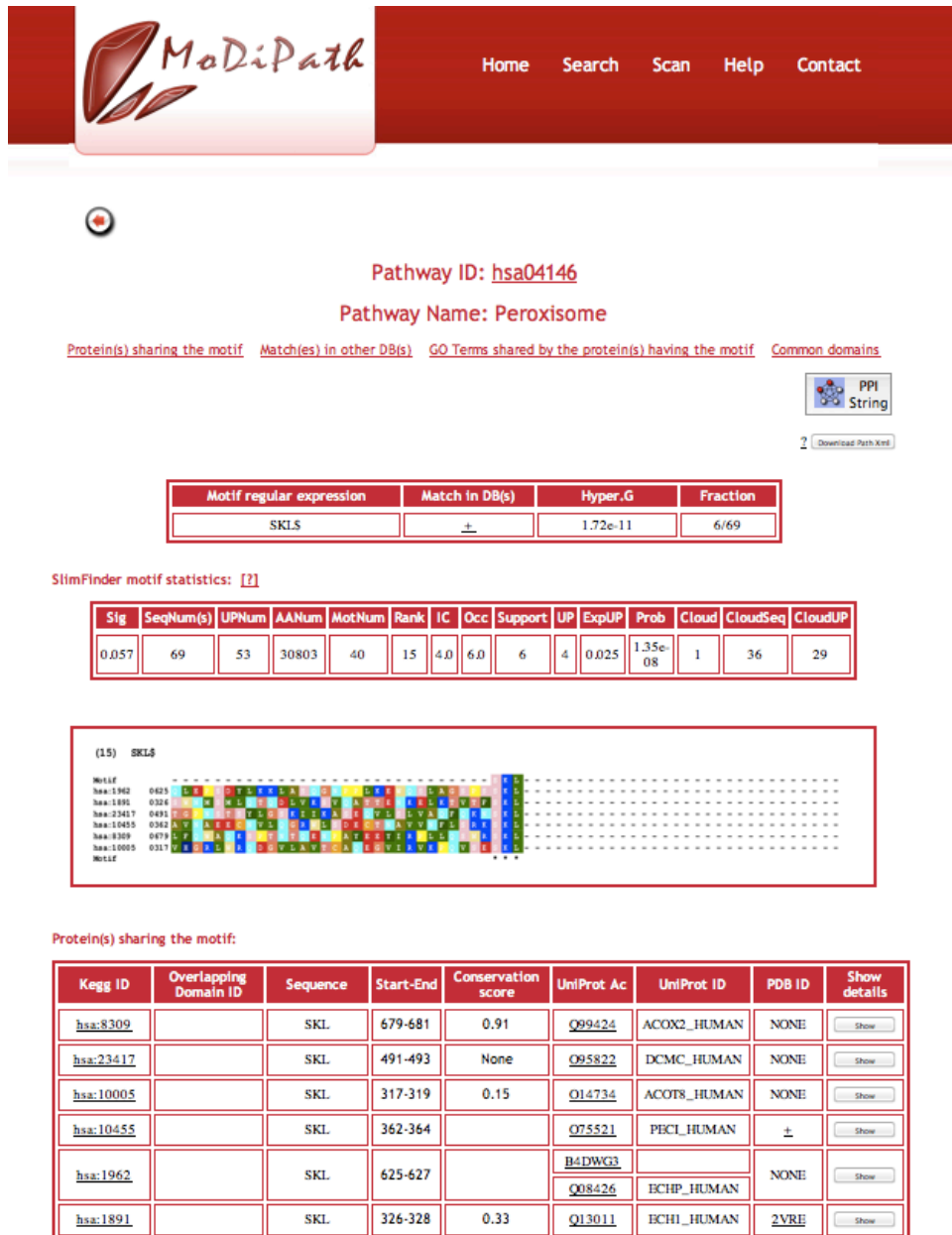
We found the “SKL\$” motif in six proteins out of the sixty-nine belonging to

the Peroxisome pathway. All of them are localized in the Peroxisome, five of them participate to the fatty acid metabolic process and three of them have a catalytic activity (Figure 10). Notably the number of proteins matching this motif by chance is very limited; apart from the 6 proteins predicted by our procedure, the motif occurs only in 8 other proteins out of the 14,239 proteins of the non-redundant UniProt human dataset (filtered at the 40% sequence identity level). We manually inspected the 8 proteins one by one, and we found that four of them are known to be membrane or secreted proteins, which means that they are likely to be false positives. The remaining four proteins are: a peroxisomal acyl-coenzyme A oxidase 3 (UniProt O15254-1), a Lon protease homolog (Q86WA8), a peroxisomal leader peptide-processing protease (Q2T9J0), and a zinc-binding alcohol dehydrogenase domain-containing protein (Q8N4Q0). O15254-1 is a different isoform of O15254-2, a human protein, reported to belong to the hsa04146 KEGG pathway that does not contain the motif and differs from O15254-1 for the lack of the last 75 C-term amino acids; it is not clear why O15254-2 was chosen for inclusion in the KEGG hsa04146 pathway; we argue that O15254-1 should be added to the KEGG hsa04146 pathway and the assignment of O15254-2 reassessed. Q86WA8 is annotated in UniProt for having the SKL\$ targeting motif and its cellular compartment is known to be the Peroxisome, but it is not associated with any KEGG pathway. Q2T9J0 and Q8N4Q0 are peroxisomal proteins but they are neither annotated for having the motif nor associated with any KEGG pathway. We propose that Q2T9J0 and Q8N4Q0 use the SKL\$ motif as targeting signal to the Peroxisome and suggest that their inclusion, and that of Q86WA8, in the KEGG Peroxisome pathway should be considered.



**Figure 9:** The Peroxisome KEGG pathway (KEGG ID: hsa04146).





**Figure 10:** MoDiPath output for the SKL\$ motif of the Peroxisome KEGG pathway (KEGG ID: hsa04146). The first and second tables report information about the motif. The third table is the multiple alignments of the proteins containing the motif. The last table lists all the proteins where the motif occurs and provides some more detailed information (e.g. position of the motif, conservation score).

### 3.1.1.2. Newly discovered motif: “[FL].L.C..Y..A”

We illustrate here the case of the [FL].L.C..Y..A motif, which we found to be significantly over-represented in the human Fc gamma R-mediated phagocytosis KEGG pathway (hsa04666).

The KEGG pathway hsa04666 plays an important role in the host defense through the phagocytosis mechanism. The opsonization with antibodies (IgG) of infectious pathogens makes them recognizable by Fc gamma receptors and more susceptible to the action of phagocytes. Fc gamma receptors trigger through several signals the phosphorylation of many proteins leading to the formation of phagosomes (Kedzierska, et al., 2001). Using our approach we discover that protein phosphorylations is probably mediated by a linear motif “[FL].L.C..Y..A” found in five proteins out of sixty-three belonging to the same pathway (hsa04666), since it was reported that they are involved in the interactions with phosphoinositides (PtdIns) (see Table 3). Phosphoinositides represent a small fraction of cellular phospholipids and are very important regulatory molecules utilized both as cellular membrane structural lipids and as precursors of multiple signaling molecules.

Uniprot ID	Protein Name	Protein Function
P42338	Phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit beta isoform	phosphorylates : PtdIns, PtdIns4P, PtdIns(4,5)P2
Q9Y217	1-phosphatidylinositol-3-phosphate 5-kinase	Regulated by PI(3,5)P2
Q13393	Phospholipase D1	Stimulated by PtdIns(4,5)P2 and PtdIns(3,4,5)P3
Q92608	Dedicator of cytokinesis protein 2 (DOCK2)	Translocation dependt to PtdIns(3,4,5)P3
O14939	Phospholipase D2	Stimulated by PtdIns(4,5)P2

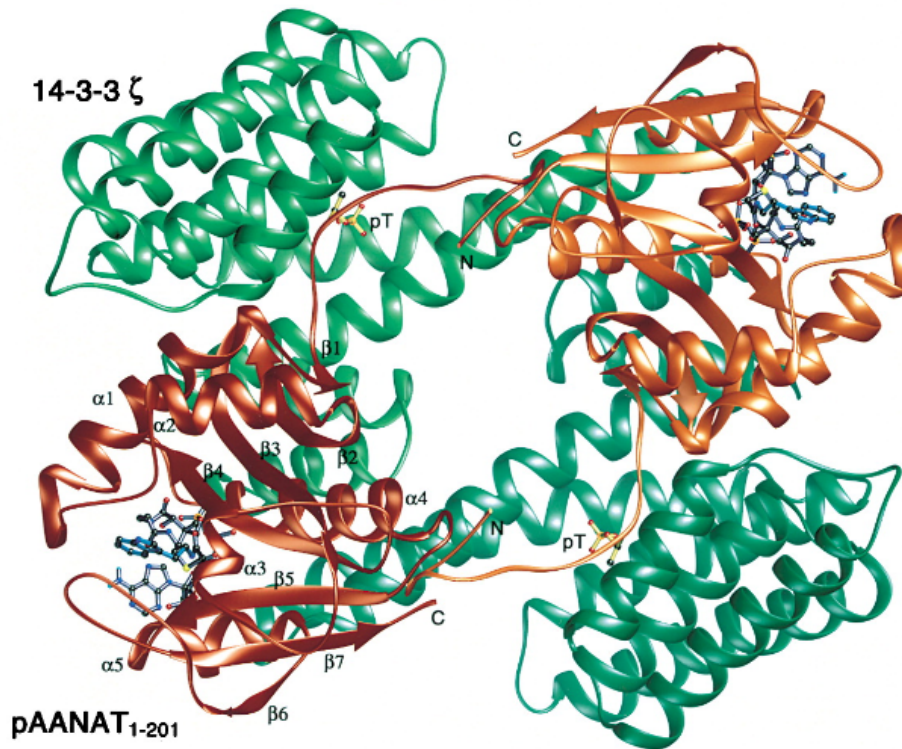
**Table 3:** List of proteins found by MoDiPath that match the motif “[FL].L.C..Y..A” and their respective functions.

By searching the motif in the whole set of human UniProt sequence, we found 9 additional occurrences in 9 different proteins. Three of them are isoforms of Q13393 and two are isoforms of O14939. Of the remaining four, one (O00329) is a PtdIns(4,5)P2 3-kinase catalytic subunit delta isoform, which is reported to be involved in the PtdIns phosphate biosynthesis, and one (Q8TDW7) is the Protocadherin FAT-3. The molecular function of FAT-3 is not well known, however some authors (Lesa, et al., 2003; Marza, et al., 2008) reported that the fat-3 gene acts in the same genetic pathway as synaptojanin, the main substrate of which in the brain is PtdIns(4,5)P2 and suggest that FAT-3 functions in the endocytic part of the synaptic vesicles recycling process. More specifically, Marza et al (Marza, et al., 2008) found that the levels of PtdIns(4,5)P2 at release sites are increased in *Caenorhabditis elegans* fat-3 mutants lacking long-chain polyunsaturated fatty acids (LC-PUFAs), which would suggest that fat-3 influences the levels of PtdIns(4,5)P2 at release sites. For the remaining two proteins (O75976 and Q8NEZ3) we did not find any clue to deduce potential interactions with phosphoinositides and we cannot exclude that they are false positives. We also analysed the 58/63 hsa04666 proteins that do not have the “[FL].L.C..Y..A” motif. In this case, we automatically selected proteins that have at least one keyword related to phosphoinositides (e.g. PtdIns) in their UniProt annotation: we found ten of such proteins and inspected their sequences. In six of them, we found motifs that are similar, although not identical, to “[FL].L.C..Y..A”. For example, the P48736 sequence contains the subsequence FVYSCAGYCVA which could be described by the “[FL].[LY].C..Y..A” regular expression, a less specific version of the original expression. In the four remaining sequences, we did not find subsequences sufficiently similar to the identified motif.

In conclusion, our analysis suggests that the “[FL].L.C..Y..A” motif (and perhaps other related ones) is involved or participates in the recognition of phosphoinositides.

## 4. Computational approach to the study of the g14-3-3 interactor network

The fine-tuning of the phosphorylation/dephosphorylation status of proteins is widely used by eukaryotic cells to regulate multiple cellular processes. In this scenario, in addition to the fundamental activity of different protein kinases and phosphatases, a key role is played by dimeric 14-3-3, a protein belonging to a highly conserved protein family, that binds to specific Ser/Thr phosphorylated sites on target proteins. The interaction of 14-3-3 with the target proteins is mediated by conserved residue located in an amphipatic groove of each monomer and requires specific binding motifs on the targets (Figure 11) (Obsil, et al., 2001). Three general consensus sequences for 14-3-3 binding have been defined so far: the mod-1 motif RS.[ST].P, the mod-2 motif R..[ST].P (Muslin, et al., 1996; Yaffe, et al., 1997), and the mode-3 motif [ST].{1-2}\$ (Coblitz, et al., 2006). Moreover, there are also proteins that interact with 14-3-3s through other phosphopeptide sequences and in some cases through non-phosphorylated sequences (Aitken, 2006; Hallberg, 2002; Petosa, et al., 1998). Specific work and large-scale proteomic studies in different organisms, from yeast to human, have led to the identification of hundreds of intracellular 14-3-3 target proteins including enzymes and structural components of metabolism (more than 200 14-3-3 target proteins in Human) (Johnson, et al., 2010).



**Figure 11:** Structure of the 14-3-3 in complex with pAANAT<sub>1-201</sub> (PDB code 1IB1) (Obsil, et al., 2001). 14-3-3 is shown in green and pAANAT<sub>1-201</sub> in brown. AANATs contain a conserved sequence motif, “RRHTLP” (residues 28–33 of ovine AANAT). The phosphorylated Thr-31 residue of pAANAT<sub>1-201</sub>, is shown in yellow, and the bisubstrate in blue.

#### 4.1. g14-3-3 interactor network

In this project we analyzed the 14-3-3 interactor network of the protozoan parasite *Giardia duodenalis* (g14-3-3). *G. duodenalis* is a flagellated protozoan that parasitizes the upper part of the small intestine of mammals causing giardiasis, the most common non-bacterial and non-viral diarrheal diseases, estimated to infect 280 million people each year (Thompson, 2000). Due to the evident role of g14-3-3s in the parasite developing processes, the identification of the 14-3-3 interacting partners would provide novel information on the biology of *Giardia*. To this aim, a MS-based

proteomic analysis of *in vivo* affinity purified g14-3-3 complexes from *Giardia* were performed at ISS (*Istituto Superiore di sanita*, Rome, Italy).

### 4.1.1. Results

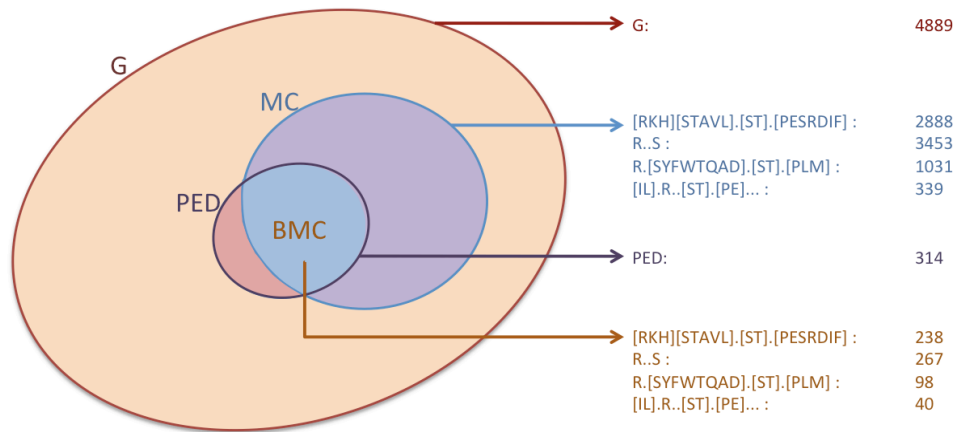
As described in paper III “The interaction network of the 14-3-3 protein in the ancient protozoan parasite *Giardia duodenalis*”, the interaction of protein targets with g14-3-3 occurs in most of the cases through well-defined phosphorylated motifs (Aitken, 2006). Three hundred fourteen (314) putative g14-3-3 protein targets were identified using a large proteomics study. Starting from these targets, we wanted to identify proteins interacting via a linear motif with g14-3-3.

#### 4.1.1.1. Sequence analysis

The interaction of g14-3-3 with its partners occurs, in most of the cases, through well-defined phosphorylated motifs (Aitken, 2006). From several databases (e.g. ELM (Puntervoll, et al., 2003), MnM (Rajasekaran, et al., 2009), Phospho-MotifFinder (Amanchy, et al., 2007)) we collected 22 regular expression encoding motifs that mediate 14-3-3 interactions in different species. We also used a g14-3-3 motif (regular expression [IL].R.[ST].[PE].[IL]) previously identified in a recent work (Lalle, et al., 2010). Fourteen out of the 23 motifs match at least one of the 314 14-3-3 putative targets.

#### 4.1.1.2. Statistical significance estimation

To statistically assess whether any of the identified motifs were significantly over-represented in the experimentally identified 14-3-3 target proteins compared to the whole *Giardia* proteome, we calculated the hyper-geometric p-value for each of the detected motifs (see legend to Figure 12).



**Figure 12:** Distribution of the proteins belonging to the Giardia genome. Giardia (G): Indicate the whole Giardia proteome. Motif containing (MC): Includes the proteins containing the motif in the whole Giardia proteome. Protein experimentally determined (PED): Represents the proteins experimentally identified as interacting with g14-3-3. Binding motif containing (BMC): Represents the intersection of the two latter protein sets.

As a result, we found that 4 out of the 14 putative 14-3-3 motifs are significantly over-represented with a hypergeometric p-value below 0.01. Details are listed in Table 4. This result highlights the existence of many true positives in the pool of proteins co-purified with g14-3-3. Only 19 out of the 314 proteins experimentally identified did not match any of the selected motifs and most of them were identified as part of the ribosomal multiprotein complex, thus suggesting that these proteins are not direct interactors of g14-3-3. Nevertheless, the presence of a motif is not sufficient to ensure that a protein interacts with g14-3-3 through that motif. In fact, gTLL3 and gDIP2, two of the experimentally identified proteins, on one hand, display the 14-3-3 interacting motif but on the other, have functions (polyglycylase and deglycylase, respectively) suggesting that their interaction with 14-3-3 occurs through a mechanism not involving the motif. To reduce the number of false positives, we tried to identify a more specific motif starting from



R..S, which is one of the 4 putative 14-3-3 motifs reported to be significantly over-represented in 267/314 proteins of our experimental data set (Table 4).

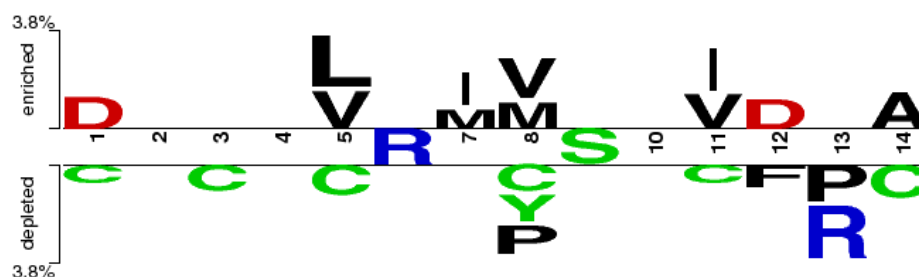
Motif Regular Expression	Fraction1: (BMC)/(PED)	Fraction2: (MC)/(G)	Hypergeometric p-value
[RKH][STAVL].[ST].[PESRDIF]	238/314	2888/4889	1,02E-10
R..S	267/314	3453/4889	6,33E-10
R.[SYFWTQAD].[ST].[PLM]	98/314	1031/4889	8,98E-06
[IL].R..[ST].[PE]...	40/314	339/4889	9,30E-05

**Table 4:** Hypergeometric p-value for the four statistically significant g14-3-3 binding motifs.

### 4.1.1.3. Protein sequence alignments

We further investigated the core motif R..S, including five upstream and five downstream residues in the experimentally identified proteins (positive sample) and compared them with those in the Giardia proteome presenting the set of Giardia proteins that match the motif but was not experimentally identified (negative sample).

In order to visualize the differences between these two groups, a WebLogo plot was performed using TwoSampleLogo (Vacic, et al., 2006). The program generates a graphical representation of statistically significant position-specific differences in amino acids between two sets of multiply aligned sequences (Figure 13). The logo is composed of three parts: (1) the upper part displays residues enriched in the positive sample; (2) the middle part displays the “R..S” motif; and (3) the lower part displays residues depleted in the positive set. Symbol height is directly proportional to residues enrichment.



**Figure 13:** Two sample logo plot representing significantly enriched (upper part) and depleted (lower part) residues in the proteins of the positive sample with respect to those in the negative sample. The “R..S” motif sequences were used as guide to build the multiple alignments.

As it can be appreciated from figure 13, in the positive set the region surrounding the motif “R..S” contains hydrophobic residues (L, I, M and V) (positions -4, -2, -1 and +2) and also charged residues (D) in positions -8 and +3. The depletion of Cysteines (C) in the motif region suggests that flexibility is required for the motif to mediate the interaction. This result is quite different from those reported in the literature about the 14-3-3 binding site, where a Leu or an Arg are observed at position -5 and a Ser or a Pro are observed at position +2 (Johnson, et al., 2010).

#### 4.1.1.4. Disorder prediction

Since it has been observed that more than 90% of the characterized 14-3-3 protein partners contain disordered regions and almost all 14-3-3 binding sites are inside disordered regions (Bustos and Iglesias, 2006; Johnson, et al., 2010) we evaluated, using IUPred (Dosztanyi, et al., 2005), the presence of disordered regions in the identified proteins. IUPred takes a protein sequence as input and gives, for each amino acid a score ranging from 0 (complete order) to 1 (complete disorder). Residues with a score above 0.5 are considered as disordered. In this study a motif is considered as disordered if all its residues have a disorder score  $> 0.5$ . An example is shown in Figure

14. The “R..S” motifs are enclosed between vertical lines (“RIPS” from 201 to 204 and “REAS” from 587 to 590). Intriguingly, almost 25% of motifs in the proteins identified experimentally were predicted to be disordered and 82% of them were predicted to be exposed. The solvent accessibility of amino acids in the protein sequence was predicted using NetSurfP (Petersen, et al., 2009) (see Material and Methods).



**Figure 14:** IUPred disorder prediction plot for the motif “R..S” on the protein sequence GL50803\_3206. Detected motifs are enclosed between vertical lines. IUPred was used to predict the disorder propensity of each amino acid of the protein sequences. The tool gives a score (y axis) ranging from 0 to 1 for each residue (x axis), with 0.5 being the suggested threshold above which a residue is considered disordered. We define a motif as being part of a disordered region if all its conserved residues have a score above 0.5 (e.g. In the motif “R..S”, R and S are the fixed positions).

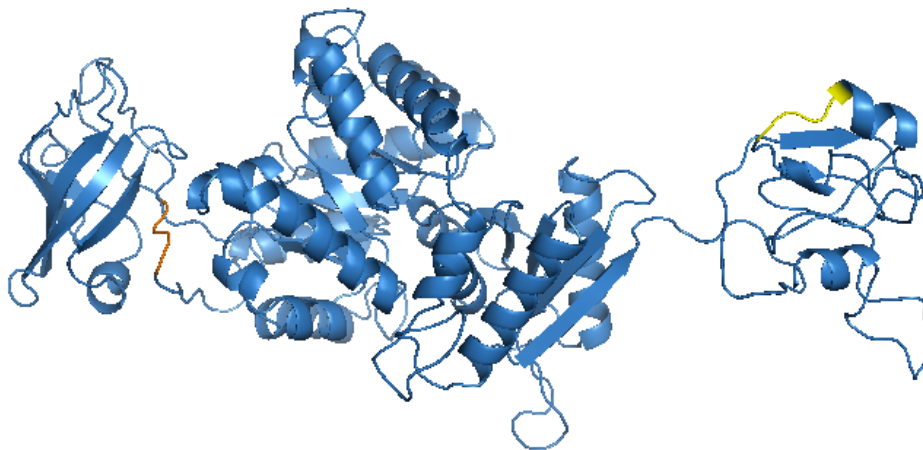
#### 4.1.1.5. Structure prediction

We further tried to predict the 14-3-3 motif solvent exposure using the tertiary structure of proteins hosting the motif. When the structure of a protein is known, solvent accessibility can be directly computed from its atomic coordinates. Only a few of our putative 14-3-3 binding proteins have a known structure. For the remaining ones, one could either predict the accessibility from the sequence alone, or if a homologue of known structure is available, build a comparative model of the structure and use it to compute approximate accessibility values.

We were able to model 87 proteins out of 314. Model quality was evaluated

using the QMEAN score (Benkert, et al., 2009). 57% of the models displayed a QMEAN score  $> 0.6$  which ranges from 0 (low quality model) to 1 (high quality model). 84 out of the 175 motifs present in the modeled proteins were found to be exposed to the solvent (average accessibility value above 25%, Solvent accessibility was calculated using POPS (Cavallo, et al., 2003)).

As an example, the Pyruvate kinase three-dimensional model (GL50803\_3206) with the localization of the putative 14-3-3 binding motif is shown in Figure 15.



**Figure 15:** 3d structure model of the protein Pyruvate kinase (GL50803\_3206). The motif 'REAS' is in yellow and the motif 'RIPS' is in orange. We used HHpred (Söding J et al., 2005) to identify homologous protein of known structure (template) and Modeller (Fiser A et al., 2003) for building the models. Parameters used were 80% coverage and E-value $<10^{-3}$ . The figure was drawn with the PyMOL program (<http://www.pymol.org/>).

## 5. Conclusions and outlook

The vast amount of functions encoded by SLiMs makes them important targets for study. With the growth of experimental data available in the last few years, SLiM discovery has become a challenging research field. SLiMs are today a primary source of protein function prediction. Novel computational methods and tools give us the possibility to better explore and analyze SLiMs. In this thesis I worked on three crucial aspects regarding SLiMs: the motif identification and annotation process (Paper I), the construction and testing of a SLiM predictor (Paper II) and the use of the information extracted from a real biological protein network to improve the specificity and sensitivity of known motifs (Paper III).

The biological role of a short linear motif can be accessed via experimental validation, high-throughput computational analyses or by carefully reviewing the literature. Several projects have collected information about SLiMs and store them in databases. In this thesis I presented the ELM resource, a manually curated database currently covering more than 159 SLiMs.

The computational discovery of linear motifs is a difficult task, which usually requires the identification of a set of non-homologous proteins sharing a common functional feature (e.g., an interaction partner or a cellular compartment). Many algorithms for motif discovery are nowadays available and appropriate statistics have been developed for estimating the effectiveness of a motif for function prediction. However, several challenging aspects still remain. For example, the identification of

appropriate sets of non-homologous proteins sharing a functional feature or the association of appropriate biological functions to newly discovered motifs are difficult tasks. To this aim we developed MoDiPath, a motif discovery tool for the identification of motifs in KEGG pathways. We were able to identify a high number of potentially biologically meaningful motifs, which represent a valid starting point for further computational and experimental functional investigation. The methodology is reliable, as demonstrated by the fact that it re-discovered many known motifs (e.g. the targeting Peroxisome signal SLK\$). Furthermore, it demonstrated to be a promising tool for the discovery of novel motifs (e.g. [FL].L.C..Y..A).

Computational methods aimed at studying and discovering motifs are today a great aid and complement for experimental studies. In this thesis we illustrated the case of proteins belonging to an experimentally determined g14-3-3 interaction network and assumed to be interacting with g14-3-3 through a linear motif. Our computational approach made it possible to reduce the number of false positives and to define a more specific motif mediating the interaction.

As a concluding remark, I believe that in the next years we will witness an increasing interest of the scientific community in functional motifs and expect advances in areas such as motif statistics, and motif discovery algorithm design. These could help enrich our understanding of the cellular biology and may play an important role in the comprehension of human diseases. We hope this work will contribute to speed up the discovery of novel motifs and will constitute a useful resource for life scientists.

## 6. Materials and Methods

In the following paragraphs, I will briefly describe some tools and methods used to develop ELM, MoDiPath and to study the g14-3-3 interaction network.

### 6.1. CD-HIT

CD-HIT (Cluster Database at High Identity with Tolerance) (<http://www.bioinformatics.org/cd-hit/>) (Li and Godzik, 2006) is a fast sequence clustering program, it groups similar proteins into clusters based on their sequence similarities. Sequences are sorted from the longest to the shortest. The longest sequence becomes the representative of the first cluster, and then the next sequences are compared one by one to the representative. If one of them is similar enough according to a certain threshold, then it is included into the cluster, if not a new cluster starts with this sequence as a representative. In the end each cluster is replaced by its representative sequence. In this study, the last release of CD-HIT 4.0 beta was used to remove redundancy from sequences at the 40% identity level. PSI-CD-HIT a sub version of CD-HIT was used to lower the threshold to 25% sequence identity.

### 6.2. SlimFinder

SLiMFinder (Short Linear Motif Finder) (<http://bioinformatics.ucd.ie/shields/software/slimfinder/>) (Edwards, et al., 2007) represents one of the most promising tools for *de novo* linear motif discovery. It is composed by two algorithms, SLiMBuild and SLiMChance. SLiMBuild tries first to identify linear motifs shared among unrelated

proteins. To do this proteins are first clustered according to their sequence composition using Blast (Altschul, et al., 1990). Then the motif is built by combining shared dimers adjacent in the sequence into longer patterns. A third step consists in adding amino acid degeneracy and/or wildcards by adding variants that occur in the unrelated proteins. Once the motif is constructed, SLiMChance calculates a score that assesses motif over-representation adjusted for evolutionary relationships. SLiMFinder also offers some input masking options to restrict the area of research to a smaller number of proteins. Masking options are disorder sequence, low complexity region and Uniprot features.

In this work the SLiMFinder release 4.0 was run locally with default parameters. Only two parameters were modified: (1) Disorder masking, the option to mask disorder region in sequences, was deactivated (`dismask=F`); (2) and the probcut cut-off was fixed to 0.99. The top  $x$  motifs within the probcut threshold were retrieved.

### 6.3. CompariMotif

CompariMotif (Motif-Motif comparison software) (<http://bioinformatics.ucd.ie/shields/software/comparimotif/>) (Edwards, et al., 2008) identifies which motifs have some degree of overlap with another motif. It starts by formatting the regular expression of the two motifs, and then it compares them in a pairwise fashion. CompariMotif takes as input two lists of motifs and compares all the possible pair combinations. As output, it provides a list of motif pairs and their score (NormalIC). This score represents the degree of similarity (information content), and it varies from weak (0.0) to strong similarity (1.0).



## 6.4. Conservation Score

Conservation Score (CS) (<http://conscore.embl.de>) (Chica, et al., 2008), a newly developed tool, uses evolutionary conservation to evaluate the functionality of a linear motif or to assess the power of a new motif regular expression. The Conservation score algorithm includes three stages. First, a set of homologous sequence is created and used to reconstruct their evolution, while conserving the motif. Next each sequence is weighted based on the observed evolutionary events. Lastly a conservation score (CS) is calculated.

Conservation score requires as input the protein sequence containing the motif and the motif itself. The CS varies from 0 to 1. A score of 1 means that the predicted motif is fully conserved in all the informative sequences, while a score of 0 means that, the predicted motif is present only in the query sequence.

## 6.5. IUPred

IUPred (Intrinsically Unstructured Prediction) (<http://iupred.enzim.hu/>) (Dosztanyi, et al., 2005) is a prediction method aimed at identifying intrinsically unstructured/disordered region in protein sequences. The method is based on the estimation of the capacity of polypeptides to form stabilizing contacts, by calculating the total pair-wise inter-residue interaction energy. The tool takes as input a protein sequence and gives as output a score for each residue ranging from 0 to 1, with 0.5 being the suggested threshold above which a residue is considered to be disordered.

## 6.6. TwoSampleLogo

TwoSampleLogo (Two Sample Logo ) (<http://www.twosamplelogo.org/>)

(Vacic, et al., 2006) is a web-based tool used to detect statistically significant differences in residues between the two sets of multiple sequence alignments, and display them as a Logo. Three sections compose the logo: (1) the upper section represents a set of over-represented residues in the positive sample; (2) the middle section displays the conserved residues of the motif; and (3) a lower section displays residues under-represented in the positive sample. In our case we have used TwoSampleLogo to find enriched residues in two groups of sequences that share a common motif but they differ in functional annotation.

### 6.7. HHPred

HHPred (<http://toolkit.tuebingen.mpg.de/hhpred>) (Soding, et al., 2005) is a powerful tool for homology detection and structural prediction. Starting with a query sequence the tool searches via PSI-Blast for similar proteins in the non-redundant database from NCBI and then builds an alignment of homologs. This alignment is used later to assign secondary structure information and to construct a HMM profile. HHpred represents the database of proteins by profile HMMs. The database of HMMs is precalculated in the same fashion using PSI-BLAST and it also contains secondary structure information. Next HHsearch a software for HMM-HMM comparison, is used to search the database of HMMs with a query HMM. The output of HHpred and HHsearch is a ranked list of database matches. Out of the ranked proteins the selected templates and the query sequence are returned as a multiple alignment in PIR format, and used by MODELLER (software implemented in HHpred) to build a 3D model (Fiser and Sali, 2003).

MODELLER (<http://salilab.org/modeller/>) is a software that builds a protein three-dimensional structure (3D) using homology or comparative modeling. The tool predicts a 3D model of a protein sequence (query) based on its

alignment to at least one related protein of known structure (template). The prediction process consists of four main steps: (1) identification of structural template(s), (2) query sequence-template structure(s) alignment, (3) model building, (4) and model quality evaluation.

### 6.8. QMEAN

QMEAN (Qualitative Model Energy ANalysis) (<http://swissmodel.expasy.org/qmean>) (Benkert, et al., 2009) is a tool for 3D model quality estimation, essential for protein structure prediction. The tool gives the possibility to evaluate a single model using a QMEAN scoring function or to compare a list of models of the same target protein all-against-all using QMEANclust scoring function. The QMEAN scoring function uses a single model to calculate local and global pre-residue quality based on the combination of different structural descriptors. QMEANclust depends essentially on the composition of the assessed set of models. It combines cluster information with single model quality estimated by QMEAN. As a result models are ranked based on the QMEAN or QMEANclust score varying from 0 to 1 reflecting model quality.

### 6.9. NetSurfP

NetSurfP (Protein Surface Accessibility and Secondary Structure Predictions) (<http://www.cbs.dtu.dk/services/NetSurfP/>) (Petersen, et al., 2009) is an artificial neural network based method that predicts surface accessibility and secondary structure of an amino acid in a polypeptide chain. The method assigns a reliability score for each residue that reflects the degree of surface exposure, usually calculated as a solvent accessible surface area (ASA) or relative surface area (RSA). Based on this calculation the residue is defined as exposed or buried, above or below 25% maximum

solvent exposure ( $ASA_{\max}$ ) respectively.

### 6.10. POPS

POPS (Parameter Optimised Surfaces) (<http://mathbio.nimr.mrc.ac.uk/wiki/POPS>) (Fraternali and Cavallo, 2002) is a fast method to calculate solvent accessibility surface areas (SASAs) from a given structure at the atomic and residue level. The method is based on a probabilistic formula fast to compute, proposed previously by Wodak and Janin (Wodak and Janin, 1980). The formula was then re-parameterised to produce accurate atomic (POPS-A) and residue (POPS-R) SASAs. As input the tool requires a structure (PDB file) and gives us output detailed atomic and residue SASAs.

## 7. References

- Aitken, A. (2006) 14-3-3 proteins: a historic overview, *Seminars in cancer biology*, **16**, 162-172.
- Alfarano, C., *et al.* (2005) The Biomolecular Interaction Network Database and related tools 2005 update, *Nucleic acids research*, **33**, D418-424.
- Altschul, S.F., *et al.* (1990) Basic local alignment search tool, *Journal of molecular biology*, **215**, 403-410.
- Amanchy, R., *et al.* (2007) A curated compendium of phosphorylation motifs, *Nature biotechnology*, **25**, 285-286.
- Bairoch, A. (1993) The PROSITE dictionary of sites and patterns in proteins, its current status, *Nucleic acids research*, **21**, 3097-3103.
- Bairoch, A. and Boeckmann, B. (1994) The SWISS-PROT protein sequence data bank: current status, *Nucleic acids research*, **22**, 3578-3580.
- Beevers, C.S., *et al.* (2006) Curcumin inhibits the mammalian target of rapamycin-mediated signaling pathways in cancer cells, *International journal of cancer. Journal international du cancer*, **119**, 757-764.
- Benkert, P., Kunzli, M. and Schwede, T. (2009) QMEAN server for protein model quality estimation, *Nucleic acids research*, **37**, W510-514.
- Berggard, T., Linse, S. and James, P. (2007) Methods for the detection and analysis of protein-protein interactions, *Proteomics*, **7**, 2833-2842.
- Beugnet, A., Wang, X. and Proud, C.G. (2003) Target of rapamycin (TOR)-signaling and RAIP motifs play distinct roles in the mammalian TOR-dependent phosphorylation of initiation factor 4E-binding protein 1, *The Journal of biological chemistry*, **278**, 40717-40722.
- Brown, N.P., Leroy, C. and Sander, C. (1998) MView: a web-compatible database search or multiple alignment viewer, *Bioinformatics*, **14**, 380-381.
- Burke, P.A., *et al.* (2002) Cilengitide targeting of alpha(v)beta(3) integrin receptor synergizes with radioimmunotherapy to increase efficacy and apoptosis in breast cancer xenografts, *Cancer research*, **62**, 4263-4272.
- Bustos, D.M. and Iglesias, A.A. (2006) Intrinsic disorder is a key characteristic in partners that bind 14-3-3 proteins, *Proteins*, **63**, 35-42.
- Carroll, M., Dyer, J. and Sossin, W.S. (2006) Serotonin increases phosphorylation of synaptic 4EBP through TOR, but eukaryotic initiation factor 4E levels do not limit somatic cap-dependent translation in aplysia neurons, *Molecular and cellular biology*, **26**,

- 8586-8598.
- Cavallo, L., Kleinjung, J. and Fraternali, F. (2003) POPS: A fast algorithm for solvent accessible surface areas at atomic and residue level, *Nucleic acids research*, **31**, 3364-3366.
- Chiang, G.G. and Abraham, R.T. (2007) Targeting the mTOR signaling network in cancer, *Trends in molecular medicine*, **13**, 433-442.
- Chica, C., *et al.* (2008) A tree-based conservation scoring method for short linear motifs in multiple alignments of protein sequences, *BMC bioinformatics*, **9**, 229.
- Coblitz, B., *et al.* (2006) C-terminal binding: an expanded repertoire and function of 14-3-3 proteins, *FEBS letters*, **580**, 1531-1535.
- Copley, R.R. (2005) The EH1 motif in metazoan transcription factors, *BMC genomics*, **6**, 169.
- Davey, N.E., Edwards, R.J. and Shields, D.C. (2010) Computational identification and analysis of protein short linear motifs, *Frontiers in bioscience : a journal and virtual library*, **15**, 801-825.
- Davey, N.E., Shields, D.C. and Edwards, R.J. (2006) SLiMDisc: short, linear motif discovery, correcting for common evolutionary descent, *Nucleic acids research*, **34**, 3546-3554.
- Davey, N.E., Shields, D.C. and Edwards, R.J. (2009) Masking residues using context-specific evolutionary conservation significantly improves short linear motif discovery, *Bioinformatics*, **25**, 443-450.
- Davey, N.E., Trave, G. and Gibson, T.J. (2011) How viruses hijack cell regulation, *Trends in biochemical sciences*, **36**, 159-169.
- Davey, N.E., *et al.* (2011) Attributes of short linear motifs, *Molecular bioSystems*.
- Diella, F., *et al.* (2004) Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins, *BMC bioinformatics*, **5**, 79.
- Diella, F., *et al.* (2008) Understanding eukaryotic linear motifs and their role in cell signaling and regulation, *Frontiers in bioscience : a journal and virtual library*, **13**, 6580-6603.
- Dosztanyi, Z., *et al.* (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content, *Bioinformatics*, **21**, 3433-3434.
- Edwards, R.J., Davey, N.E. and Shields, D.C. (2007) SLiMFinder: a probabilistic method for identifying over-represented, convergently evolved, short linear motifs in proteins, *PloS one*, **2**, e967.
- Edwards, R.J., Davey, N.E. and Shields, D.C. (2008) CompariMotif: quick and easy comparisons of sequence motifs, *Bioinformatics*, **24**, 1307-1309.

- Finn, R.D., *et al.* (2010) The Pfam protein families database, *Nucleic acids research*, **38**, D211-222.
- Fiser, A. and Sali, A. (2003) Modeller: generation and refinement of homology-based protein structure models, *Methods in enzymology*, **374**, 461-491.
- Fraternali, F. and Cavallo, L. (2002) Parameter optimized surfaces (POPS): analysis of key interactions and conformational changes in the ribosome, *Nucleic acids research*, **30**, 2950-2960.
- Gabaldon, T. (2010) Peroxisome diversity and evolution, *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **365**, 765-773.
- Gould, C.M., *et al.* (2010) ELM: the status of the 2010 eukaryotic linear motif resource, *Nucleic acids research*, **38**, D167-180.
- Gutman, R., *et al.* (2005) QuasiMotifFinder: protein annotation by searching for evolutionarily conserved motif-like patterns, *Nucleic acids research*, **33**, W255-261.
- Hallberg, B. (2002) Exoenzyme S binds its cofactor 14-3-3 through a non-phosphorylated motif, *Biochemical Society transactions*, **30**, 401-405.
- Hay, N. and Sonenberg, N. (2004) Upstream and downstream of mTOR, *Genes & development*, **18**, 1926-1945.
- hunt, T. (1990) Protein sequence motifs involved in recognition and targeting: a new series, *Trends in biochemical sciences*, **15**, 305.
- Igarashi, Y., *et al.* (2007) CutDB: a proteolytic event database, *Nucleic acids research*, **35**, D546-549.
- Jensen, L.J., *et al.* (2009) STRING 8--a global view on proteins and their functional interactions in 630 organisms, *Nucleic acids research*, **37**, D412-416.
- Johnson, C., *et al.* (2010) Bioinformatic and experimental survey of 14-3-3-binding sites, *The Biochemical journal*, **427**, 69-78.
- Jones, S., *et al.* (1998) Domain assignment for protein structures using a consensus approach: characterization and analysis, *Protein science : a publication of the Protein Society*, **7**, 233-242.
- Kadaveru, K., Vyas, J. and Schiller, M.R. (2008) Viral infection and human disease--insights from minimotifs, *Frontiers in bioscience : a journal and virtual library*, **13**, 6455-6471.
- Kanehisa, M. and Goto, S. (2000) KEGG: kyoto encyclopedia of genes and genomes, *Nucleic acids research*, **28**, 27-30.
- Kedzierska, K., *et al.* (2001) FcgammaR-mediated phagocytosis by human macrophages involves Hck, Syk, and Pyk2 and is augmented by GM-CSF, *Journal of leukocyte biology*, **70**, 322-328.
- Lalle, M., *et al.* (2010) Involvement of 14-3-3 protein post-translational

- modifications in *Giardia duodenalis* encystation, *International journal for parasitology*, **40**, 201-213.
- Lesca, G.M., *et al.* (2003) Long chain polyunsaturated fatty acids are required for efficient neurotransmission in *C. elegans*, *Journal of cell science*, **116**, 4965-4975.
- Letunic, I., Doerks, T. and Bork, P. (2009) SMART 6: recent updates and new developments, *Nucleic acids research*, **37**, D229-232.
- Li, W. and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, *Bioinformatics*, **22**, 1658-1659.
- Linding, R., *et al.* (2007) Systematic discovery of in vivo phosphorylation networks, *Cell*, **129**, 1415-1426.
- Marcatili, P., Bussotti, G. and Tramontano, A. (2008) The MoVIN server for the analysis of protein interaction networks, *BMC bioinformatics*, **9 Suppl 2**, S11.
- Marza, E., *et al.* (2008) Polyunsaturated fatty acids influence synaptojanin localization to regulate synaptic vesicle recycling, *Molecular biology of the cell*, **19**, 833-842.
- Massenet, C., *et al.* (2005) Effects of p47phox C terminus phosphorylations on binding interactions with p40phox and p67phox. Structural and functional comparison of p40phox and p67phox SH3 domains, *The Journal of biological chemistry*, **280**, 13752-13761.
- Michael, S., *et al.* (2008) Discovery of candidate KEN-box motifs using cell cycle keyword enrichment combined with native disorder prediction and motif conservation, *Bioinformatics*, **24**, 453-457.
- Muslin, A.J., *et al.* (1996) Interaction of 14-3-3 with signaling proteins is mediated by the recognition of phosphoserine, *Cell*, **84**, 889-897.
- Neduva, V. and Russell, R.B. (2006) DILIMOT: discovery of linear motifs in proteins, *Nucleic acids research*, **34**, W350-355.
- Neduva, V. and Russell, R.B. (2006) Peptides mediating interaction networks: new leads at last, *Current opinion in biotechnology*, **17**, 465-471.
- Obsil, T., *et al.* (2001) Crystal structure of the 14-3-3zeta:serotonin N-acetyltransferase complex. a role for scaffolding in enzyme regulation, *Cell*, **105**, 257-267.
- Ogata, H., *et al.* (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes, *Nucleic acids research*, **27**, 29-34.
- Pawson, T. and Nash, P. (2003) Assembly of cell regulatory systems through protein interaction domains, *Science*, **300**, 445-452.
- Petersen, B., *et al.* (2009) A generic method for assignment of reliability scores applied to solvent accessibility predictions, *BMC structural*



- biology*, **9**, 51.
- Petosa, C., *et al.* (1998) 14-3-3zeta binds a phosphorylated Raf peptide and an unphosphorylated peptide via its conserved amphipathic groove, *The Journal of biological chemistry*, **273**, 16305-16310.
- Plewczynski, D., *et al.* (2005) AutoMotif server: prediction of single residue post-translational modifications in proteins, *Bioinformatics*, **21**, 2525-2527.
- Puntervoll, P., *et al.* (2003) ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins, *Nucleic acids research*, **31**, 3625-3630.
- Rajasekaran, S., *et al.* (2009) Minimotif miner 2nd release: a database and web system for motif search, *Nucleic acids research*, **37**, D185-190.
- Ramu, C. (2003) SIRW: A web server for the Simple Indexing and Retrieval System that combines sequence motif searches with keyword searches, *Nucleic acids research*, **31**, 3771-3774.
- Rawlings, N.D., *et al.* (2008) MEROPS: the peptidase database, *Nucleic acids research*, **36**, D320-325.
- Richardson, J.S. (1981) The anatomy and taxonomy of protein structure, *Advances in protein chemistry*, **34**, 167-339.
- Romero, P., *et al.* (2001) Sequence complexity of disordered protein, *Proteins*, **42**, 38-48.
- Russell, R.B. and Gibson, T.J. (2008) A careful disorderliness in the proteome: sites for interaction and targets for future therapies, *FEBS letters*, **582**, 1271-1275.
- Saleem, R.A., Smith, J.J. and Aitchison, J.D. (2006) Proteomics of the peroxisome, *Biochimica et biophysica acta*, **1763**, 1541-1551.
- Sanger, F. and Coulson, A.R. (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase, *Journal of molecular biology*, **94**, 441-448.
- Sayadi, A., *et al.* (2011) Exploiting publicly available biological and biochemical information for the discovery of novel short linear motifs, *PloS one*, **6**, e22270.
- Schalm, S.S., *et al.* (2003) TOS motif-mediated raptor binding regulates 4E-BP1 multisite phosphorylation and function, *Current biology : CB*, **13**, 797-806.
- Soding, J., Biegert, A. and Lupas, A.N. (2005) The HHpred interactive server for protein homology detection and structure prediction, *Nucleic acids research*, **33**, W244-248.
- Stretton, A.O. (2002) The first sequence. Fred Sanger and insulin, *Genetics*, **162**, 527-532.
- Thompson, R.C. (2000) Giardiasis as a re-emerging infectious disease and its

- zoonotic potential, *International journal for parasitology*, **30**, 1259-1267.
- Vacic, V., Iakoucheva, L.M. and Radivojac, P. (2006) Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments, *Bioinformatics*, **22**, 1536-1537.
- Via, A., *et al.* (2009) A structure filter for the Eukaryotic Linear Motif Resource, *BMC bioinformatics*, **10**, 351.
- von Mering, C., *et al.* (2002) Comparative assessment of large-scale data sets of protein-protein interactions, *Nature*, **417**, 399-403.
- Vyas, J., *et al.* (2010) MimoSA: a system for minimotif annotation, *BMC bioinformatics*, **11**, 328.
- Waterhouse, A.M., *et al.* (2009) Jalview Version 2--a multiple sequence alignment editor and analysis workbench, *Bioinformatics*, **25**, 1189-1191.
- Wendland, M. and Subramani, S. (1993) Presence of cytoplasmic factors functional in peroxisomal protein import implicates organelle-associated defects in several human peroxisomal disorders, *The Journal of clinical investigation*, **92**, 2462-2468.
- Wodak, S.J. and Janin, J. (1980) Analytical approximation to the accessible surface area of proteins, *Proceedings of the National Academy of Sciences of the United States of America*, **77**, 1736-1740.
- Yaffe, M.B., *et al.* (2001) A motif-based profile scanning approach for genome-wide prediction of signaling pathways, *Nature biotechnology*, **19**, 348-353.
- Yaffe, M.B., *et al.* (1997) The structural basis for 14-3-3:phosphopeptide binding specificity, *Cell*, **91**, 961-971.
- Yellaboina, S., *et al.* (2011) DOMINE: a comprehensive collection of known and predicted domain-domain interactions, *Nucleic acids research*, **39**, D730-735.
- Zanzoni, A., *et al.* (2002) MINT: a Molecular INTERaction database, *FEBS letters*, **513**, 135-140.

## 8. Publications

**Paper I: ELM: the status of the 2010 eukaryotic linear motif resource**

## ELM: the status of the 2010 eukaryotic linear motif resource

Cathryn M. Gould<sup>1</sup>, Francesca Diella<sup>1</sup>, Allegra Via<sup>2</sup>, Pål Puntervoll<sup>3</sup>, Christine Gemünd<sup>1</sup>, Sophie Chabanis-Davidson<sup>1</sup>, Sushama Michael<sup>1</sup>, Ahmed Sayadi<sup>2</sup>, Jan Christian Bryne<sup>3,4</sup>, Claudia Chica<sup>1</sup>, Markus Seiler<sup>1</sup>, Norman E. Davey<sup>1</sup>, Niall Haslam<sup>1</sup>, Robert J. Weatheritt<sup>1</sup>, Aidan Budd<sup>1</sup>, Tim Hughes<sup>5</sup>, Jakub Paś<sup>6</sup>, Leszek Rychlewski<sup>6</sup>, Gilles Travé<sup>7</sup>, Rein Aasland<sup>5</sup>, Manuela Helmer-Citterich<sup>8</sup>, Rune Linding<sup>9</sup> and Toby J. Gibson<sup>1,\*</sup>

<sup>1</sup>Structural and Computational Biology Unit, European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany, <sup>2</sup>Biocomputing Group, Department of Biochemical Sciences, 'A. Rossi-Fanelli', Sapienza Università di Roma, P.le Aldo Moro, 5, 00185 Rome, Italy, <sup>3</sup>Computational Biology Unit, Bergen Centre for Computational Science, Høyteknologisenteret, Thormøhlensgate 55, <sup>4</sup>Sars Centre for Marine Molecular Biology, University of Bergen, 5008 Bergen, <sup>5</sup>Department of Molecular Biology, University of Bergen, HIB, Thormøhlensgt. 55, 5020 Bergen, Norway, <sup>6</sup>BioInfoBank Institute, Limanowskiego 24A16 60-744, Poznań, Poland, <sup>7</sup>ESBS, 1, Bld Sébastien Brandt, BP10413, 67412 Illkirch, France, <sup>8</sup>Centre for Molecular Bioinformatics, Department of Biology, University of Rome 'Tor Vergata', Via della Ricerca Scientifica, 00133 Rome, Italy and <sup>9</sup>Cellular & Molecular Logic Team, The Institute of Cancer Research (ICR), Section of Cell and Molecular Biology, SW3 6JB London, UK

Received September 14, 2009; Revised October 16, 2009; Accepted October 19, 2009

### ABSTRACT

Linear motifs are short segments of multidomain proteins that provide regulatory functions independently of protein tertiary structure. Much of intracellular signalling passes through protein modifications at linear motifs. Many thousands of linear motif instances, most notably phosphorylation sites, have now been reported. Although clearly very abundant, linear motifs are difficult to predict *de novo* in protein sequences due to the difficulty of obtaining robust statistical assessments. The ELM resource at <http://elm.eu.org/> provides an expanding knowledge base, currently covering 146 known motifs, with annotation that includes >1300 experimentally reported instances. ELM is also an exploratory tool for suggesting new candidates of known linear motifs in proteins of interest. Information about protein domains, protein structure and native disorder, cellular and taxonomic contexts is used to reduce or deprecate false positive matches. Results are graphically displayed in a 'Bar Code' format, which also displays known

instances from homologous proteins through a novel 'Instance Mapper' protocol based on PHI-BLAST. ELM server output provides links to the ELM annotation as well as to a number of remote resources. Using the links, researchers can explore the motifs, proteins, complex structures and associated literature to evaluate whether candidate motifs might be worth experimental investigation.

### INTRODUCTION

Linear motifs (LMs) are short elements embedded within larger protein sequence segments that operate as sites of regulation (1–5). They can be found in telomeric proteins (6), in proteins of the extracellular matrix (7)—and seemingly every macromolecular complex in between. Many are post-translationally modified, but not all. The essence of their function is embodied in the linear amino acid sequence and is not dependent on the tertiary structural context. Nevertheless, as a consequence of low affinity binary binding interactions, they usually act in a concerted and cooperative manner, enabling regulatory decisions to be made on the basis of multiple inputs (8–12). These properties may be important for

\*To whom correspondence should be addressed. Tel: +49 6221 387398; Fax: +49 6221 387517; Email: [gibson@embl-heidelberg.de](mailto:gibson@embl-heidelberg.de)  
Present address:

Christine Gemünd, Cellzome AG, Meyerhofstrasse 1, 69117 Heidelberg, Germany.

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

© The Author(s) 2009. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/2.5/uk/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

D168 *Nucleic Acids Research*, 2010, Vol. 38, Database issue

the inherent robustness of cellular systems (13), as cell regulation is increasingly revealed to be cooperative, networked and redundant in nature (14–20).

Over the time that we have worked to develop the Eukaryotic Linear Motif resource ELM, our conviction has grown that there will be well over a million LM instances in a higher eukaryotic proteome. (Phosphoproteomics is on the way to revealing  $\gg 100\,000$  phosphorylation sites, for example.) If these estimates reflect reality, one might expect that experimentalists should be stumbling across new motifs with every experiment. But they are not. The paradox is that it remains difficult to establish the existence of LM instances whether by experiment or computationally. The bioinformatics problem is simple to state: LMs are too short (and the information content too poor) to be statistically significant in protein sequence searches. Experimentalists are similarly afflicted: while trying to identify LMs, they are likely to spend a lot of resources, time and effort performing experiments on the false motif candidates, which usually vastly outnumber the genuine ones in any set of proteins of interest (1).

Nevertheless, useful advances are now being made in the bioinformatics tools that address the remarkable modularity of eukaryotic regulatory proteins. Thus, two dedicated LM databases now exist: ELM (21) and the Minimotif Miner (22). (Users should utilize both resources as there are many differences in approach and the datasets only partially overlap.) Specialized databases for phosphorylation sites include PhosphoSite, Phospho.ELM and Phosida (23–25). Resources such as HPRD (26) and UniProtKB/Swiss-Prot (27) annotate a broader range of Post-Translational Modifications (PTMs). Furthermore, numerous predictive tools for identifying natively disordered protein segments—the main harbour for LMs (28–30)—have become available (31,32), complementing the more established globular domain resources Pfam, SMART, PROSITE and InterPro (33–36). The ELM datasets have been used by bioinformaticians to develop and benchmark novel prediction strategies such as hunting for motifs in interaction data and to provide likelihood estimates for motif candidates based on structural and sequence conservation contexts (37–41). While LM discovery remains challenging, if progress continues apace, it should become possible to address the intricate subfunctionalization of proteins like p53, CBP/p300, APC and Tau with ever-greater effectiveness.

Here, we provide an overview of the current status of the ELM resource and the research contexts in which it is being used. The utility of ELM is threefold: for researchers, it is first a knowledgebase, second a predictive tool but ELM has a third important function too; it can also be used for more general educational purposes, as it covers a topic that is often poorly served in text books. ELM provides written text summaries and links to the experimental literature that are a useful starting point for people who, for any reason, wish to gain an understanding of the role of LMs in cell regulation. We also take the opportunity here to provide a summary of progress made by the pioneering community of bioinformatics teams that are applying ELM to develop

new tools for LM discovery. Finally, we provide some guidance about good practice and warnings about pitfalls for researchers seeking to apply ELM in experimental motif discovery.

#### WHAT ARE LMS?

To use ELM effectively, a user will need to grasp why such a resource is needed. The earliest definition of LM known to us was written in 1990 by Tim Hunt to introduce the new Protein Sequence Motifs column in *Trends in Biological Sciences* (42).

The sequences of many proteins contain short, conserved motifs that are involved in recognition and targeting activities, often separate from other functional properties of the molecule in which they occur. These motifs are linear, in the sense that three-dimensional organization is not required to bring distant segments of the molecule together to make the recognizable unit. The conservation of these motifs varies: some are highly conserved while others, for example, allow substitutions that retain only a certain pattern of charge across the motif.

This definition was written at a time when it was becoming apparent that many cellular proteins would have complex multidomain architectures and the first LMs such as KDEL, NLS, the Destruction Box of cyclin B and the fascinating KFERQ starvation-dependent lysosomal targeting motif were being reported (43–46). The definition has stood the test of time and can still serve very well today.

Sequence motifs contributing to the tertiary structure and primary function of globular domains are excluded by the definition of LM. An LM is effectively an irreducible unit of structure and function. Although LMs may be found in exposed parts of globular folds, they must be able to function independently to fit the definition: conversely, the globular domain would still have the same function if the LM was inactivated, although of course that domain function might well be dysregulated in the absence of the motif. The need to separate motif/domain functions applies to methods that seek to define new motifs. Historically, it has been difficult to develop computational methods that can distinguish short conserved segments of protein domains from LMs. Failure to make the distinction is likely to lead to false LM assignment (1), as has often happened for the nuclear export sequence (NES) as discussed by Hantschel *et al.* and Kadlec *et al.* (47,48).

Over the last few years, it has become increasingly clear that most LMs do not reside inside globular domains but instead are present in segments of natively disordered polypeptide. Often many LMs are clustered within one segment of native disorder. LMs quite frequently overlap, providing the potential for switch-like mutually exclusive functionality. For example, overlapping peptides from p53 are present in solved structures of several different protein complexes (20). Therefore, an overview of the types and locations of protein architecture modules existing in regulatory proteins provides an essential adjunct to LM investigation.



**ELM RESOURCE ARCHITECTURE**

At the core of the ELM resource is a PostgreSQL relational database with 69 tables storing data about LMs. Not all of this complexity is fully utilized: it anticipates current and future filtering strategies as well as information retrieval by users. The key information content is summarized in Figure 1. Users should make sure they grasp the importance of the three fundamental nodes in the hierarchy: the top level 'Functional Site' links to 'ELM Motif' which includes 'ELM Instances'. The top level of 'Functional Site' is essentially a biological designation with general information: for example, 'Nuclear export signal'. The 'ELM Motif' is given a more specific description, links to information pertaining to the given LM, including key literature and Gene Ontology (GO) assignments, and includes the Regular Expression pattern representing the motif: see, for example, the NES entry at [http://elm.eu.org/elmPages/TRG\\_NES\\_CRM1\\_1.html](http://elm.eu.org/elmPages/TRG_NES_CRM1_1.html). Of note, ELM is effectively motif-centric—if a regular expression cannot be defined, there is no entry in ELM. An 'ELM Instance' embodies the specific information for a motif match in a protein sequence: for example, click on the links for the NES instance in MAPKAPK2. The instances provide the essential information that supports the ELM hierarchy. Instance-containing sequences are mapped to their respective UniProt entries. A well-annotated instance may also have links to the experimental literature, the types of experiments undertaken and to informative structure entries in the PDB (49). Importantly, an instance may have a reliability value assigned by the curator: many false positive motifs have been claimed in the literature. (Note: some of the older ELM entries do not yet have well-annotated instances).

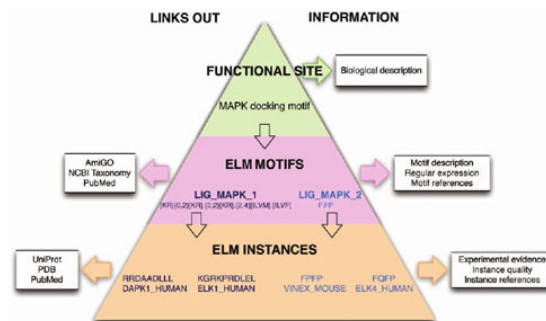
All data input is by manual curation. Annotating each ELM entry typically involves extensive literature searches,

BLAST runs, multiple alignment of relevant protein families, perusal of Swiss-Prot and other online databases and, where practical, discussion with experimentalist experts from the field. In order to promote interoperability with other bioinformatics resources, we use two public annotation standards. GO identifiers are used for cell compartment, molecular function and biological process (50) while the NCBI taxonomy database identifiers (51) are used for taxonomic nodes at the apex of phylogenetic groupings in which an LM occurs. A third standard—POSIX regular expressions (<http://standards.ieee.org/regauth/posix/>)—is used to represent the motif patterns. These 'RegExps' are conveniently usable in the Python and Perl scripting languages. They are analogous to PROSITE motifs (35), but with a different syntax. For example, the C-terminal motif `LIG_CAP-Gly_1` that binds to CAP-Gly domains for microtubule plus-end regulation (52) is represented by the RegExp

`[ed].[0,2][ed].[0,2][edq].[0,1][YF]S`

where \$ is the protein C-terminus, preceded by a conserved aromatic residue and a flexibly spaced run of negatively charged residues. See the help page [http://elm.eu.org/help.html#regular\\_expressions](http://elm.eu.org/help.html#regular_expressions) for guidance on the ELM expressions.

Table 1 provides some representative examples of different motif categories. Based on the type of function of the LM, we have defined four classes of ELM motif (Cleavage, Ligand, Modification and Target), which are summarized in the table. Some of these motifs have complicated regular expressions, others are very simple, e.g. with just two conserved positions. It has become clear that the most common conservation pattern is for three (semi-) conserved positions in the motif. A substantial minority of motifs have one or more positions that tolerate gaps



**Figure 1.** The ELM Resource hierarchy represented as a pyramid. 'Functional Site' provides a general description of the biology, for example, MAP Kinases have a docking motif in their substrates. There are more than one class of MAPK docking motifs and ELM currently provides two 'ELM Motif' entries. These contain the motif regular expression and are annotated with more specific information as well as linking out to remote resources including PubMed, NCBI Taxonomy and GO. At the base of the pyramid are the 'ELM Instances' that belong to a given 'ELM Motif' entry. The instances are annotated with information about experimental methods and instance quality and link to external resources including UniProt, PDB and PubMed.

Table 1. The four classes of LM in the ELM classification and some representative examples

Class	Class description	ELM_ID	Regular expression <sup>a</sup>	ELM description
LIG	Motifs acting as ligands to globular protein domains.	LIG_MAPK_1 LIG_APC_Dbox_1	[KR][0-2][KR];[0-2][KR];{2-4}[LVM];[LVF] R..L..[LVM]	MAPK interacting molecules (e.g. MAPKs, substrates, phosphatases) carry docking motifs that help to regulate specific interactions in the MAPK signaling networks. The classic motif approximates (R/KxxxxxxR where # is a hydrophobic residue). An KKKL-based motif that binds to the Cdk1 and Cdk20 components of APC <sup>2</sup> targets the protein for destruction in a cell cycle dependent manner.
TRG	Motifs within proteins that are sufficient for recognition and targeting to subcellular compartments.	TRG_AP2beta_CARGO_1 TRG_PEX_1	[DE];{1,2}[P][P][F][P][P][P] W...[FY]	AP2 beta appendage platform subdomain (top surface) binding motif used in targeting cargo for internalization. Specific ELM present in Pex5p and binding to Pex13p and Pex14p. Part of the peroxisomal matrix protein import system
MOD	Sites of post-translational modification of proteins.	MOD_N-GLC_1	(N)[P][ST].	Generic motif for N-glycosylation at Asparagine residues. Extracellular proteins are glycosylated in the Endoplasmic Reticulum. The first step of the process, attachment of the carbohydrate precursor, is coupled to translation and import of the nascent polypeptide, preceding folding of the protein.
CLV	Cleavage sites recognized by proteases for the processing of precursor proteins into biologically active products.	MOD_ProDKin_1 CLV_TASPASE1 CLV_PCSK_FUR_1	...(ST)P... QMLV[DG].[DE] R[RK]R.	Protein-Directed Kinase (e.g. MAPK) phosphorylation site in higher eukaryotes. Taspase is the aspartic protease which was first identified as the protease responsible for processing the trithorax (MLL) type of histone methyltransferase. Furin (PACE) cleavage site (Arg-Xaa-[Arg/Lys]-Arg;-Xaa)

<sup>a</sup>Regular expression help is available at: [http://elm.eu.org/help.html#regular\\_expressions](http://elm.eu.org/help.html#regular_expressions).

(indels). The length range of indels can usually be accurately determined from sequence alignments; the most common indel is to allow a one-residue insertion.

Table 2 provides a summary of the data that have so far been entered into the ELM DB in its current state. The most noteworthy numbers are 146 ELM motifs, the >1300 instances and the >1100 citations of LM literature. Our goal is to create representative, not comprehensive, LM entries. For abundant motifs like the sumoylation site, with thousands of instances per proteome, we will not try to annotate more than a small fraction of experimental instances, since the appropriate location for these data are the protein annotation resources such as Swiss-Prot and HPRD.

ELM is primarily developed and deployed with open source software and is hosted on CentOS Linux. Pipeline software is mainly developed in Python including some modules from the <http://BioPython.org> project to retrieve information from SWISS-PROT and PubMed. The web interface software uses the CGIModel framework (53). The server output is HTML and Javascript.

### WHY USE REGULAR EXPRESSIONS IN ELM?

The three most commonly used methods for bioinformatical representation of sequence conservation patterns are: Profile/HMMs (54); Artificial neural networks (ANNs) (55); and RegExps ([http://en.wikipedia.org/wiki/Regular\\_expression](http://en.wikipedia.org/wiki/Regular_expression)). Of these, RegExps are considered the worst approach to encapture protein sequence information. They are *ad hoc*—typically created by annotators without applying a consistent formalism. The motif characters are represented with integer values, so RegExps cannot use position-weighting to capture weaker preferences. They are over-determined and can only capture exactly what is specified (whereas the more probabilistic HMMs and ANNs can rank near misses too). They do not support searching for an exact number of a given amino acid character within a specified range [which would better approximate the charged runs in e.g. CAP-Gly and NLS motifs (56)]. Despite these shortcomings, using RegExps to establish ELM has proved to be the correct decision. Many LMs have short indels in the pattern. HMM software does not (yet) provide for variable gaps with exactly bounded ranges while ANNs do not account for gaps at all: a motif such as the NES with multiple short indels is hard to represent with these algorithms. The scoring of presence/absence matches for LM RegExps simplifies statistical analyses of motif searches. These two advantages have been critical to the first wave of development of motif-hunting software.

Thus we consider that it was appropriate to initiate LM database resources with RegExps. Of course, HMMs and ANNs are used in a number of useful predictive tools, e.g. Scansite (57) and NetPhorest (58) and there is little doubt that HMMs, neural networks and other methods will grow in importance for LM analyses in future, once the contexts can be better controlled.



Table 2. Summary of the data stored in the ELM RDB

	Number of functional site entries	ELM motifs	Instances	Links to PDB structure entries	Go terms	PubMed links	
Totals	110	146	1327	100		1125	
By category		LIG 89 MOD 30 TRG 19 CLV 8	Human 828 Mouse 104 Rat 65 Fly 47 Yeast 88 Other 195		Biological process Cell compartment Molecular function	152 69 87	From ELM motif From instance 683

### ACCESSING ELM

The ELM resource is freely accessible to users. The data in ELM can be accessed via the Web either interactively or programmatically. Motif entries are available to be browsed from the browse links page at <http://elm.eu.org/>. Details from the browse page for the LIG\_CAP-Gly\_1 entry are shown in Figure 2. A user can also submit a protein sequence of interest through the main submission page and will receive an output page with the matched candidates. The key data retrieved by the ELM resource for the sequence is displayed in a 'bar code' style graphical output as shown for the motif-rich endocytic protein Epsin-1 (Figure 3). Mouse-over provides annotation and there are many links to summaries in tabular and text form. Help is available online to explain the meanings of the elements and colour code in the output.

Programmatic access takes advantage of SOAP/XML Web Services (WS) interfaces for six ELM resource modules listed in Table 3. [See the EMBRACE registry for a large collection of Bioinformatics WS (59)]. Programmers can use the ELM DB WS interfaces to collect data—for example, a query might be to retrieve all regular expressions stored in ELM or another query might be for all ELM instances, or a defined subset thereof. Other WS interfaces allow LM matching to a query sequence and structural and conservation filtering.

Upon request, we can provide a SQL dump if for any reason, the WS interface is not suitable. At some future point, we would like to provide a standardized ELM DB dump, probably using the BioMart format (60).

### THE ELM RESOURCE FILTERS

Searches of sequence databases with short motifs do not yield significant results (due to the large number of non-functional sequences matching the motif consensus) and therefore, it is necessary to evaluate the context of the match. Essentially, any aspect of a protein that can be informative might provide contextual filtering. Filters might be simple or complicated and ELM provides examples of both. Originally, three simple filters (21) were implemented in ELM: (i) Cell compartment filter: an LM is only meaningful in appropriate cell compartments; (ii) Taxonomy filter: an LM is only meaningful in an organism that is known to possess its

interaction partners; and (iii) SMART globular domain filter: LMs are interaction sites and must be accessible, hence they are much more common in natively disordered sequence. ELM does not provide benchmarked scores for the simple filters. Two more complicated filters have been implemented and benchmarked to provide reliability assessments, for structural context and evolutionary conservation.

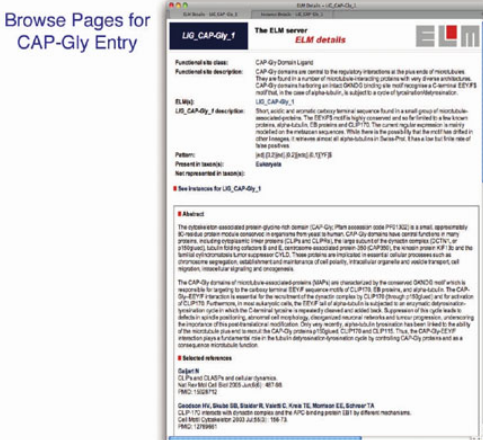
The ELM structure filter (SF) assesses the accessibility and secondary structure components of LM candidates whenever a reference globular domain structure is available (41). The benchmarked scale shows that most LMs are in exposed and accessible loops. Although a few genuine LMs are quite inaccessible in the available structural conformation, the benchmarking indicates that it is usually not worth experimental testing of the inaccessible motifs unless there is an indication of, for example, allosteric rearrangement that might enable the site to become exposed. When it applies, the SF is much more informative than the simple globular domain filter. The SF is implemented in the ELM resource output (Figure 3), and can be accessed independently as a web service (Table 3).

The ELM conservation score (CS) filter assesses the conservation of motif candidates in related proteins (61). LMs tend to be more evolutionarily dynamic than globular domains—it is uncommon to find an LM instance that is conserved between yeast and mammals (e.g. see the GLEBS and FFAT motif entries for counter-examples). The CS filter is a pipeline to collect and align homologous sequences and test ELM motifs for conservation, using a benchmarked scoring scheme. The CS filter has already proven its value in motif discovery efforts (62,63) but, due to the resource reengineering required, is not yet implemented in the ELM output. For the time being, therefore, it is offered as a stand-alone server (<http://elm.eu.org/conscorer>) and web service (Table 3). Figure 4 shows variation in conservation of some of the motif matches from the Epsin-1 example used above (Figure 3).

### THE ELM INSTANCE MAPPER

It is not uncommon that all the experimentation demonstrating the existence of a particular LM instance has been undertaken in a single model organism, e.g.

Browse Pages for CAP-Gly Entry



Functional sites:  
CAP-Gly Domain-Ligand

Functional description:  
CAP-Gly domains are found in the regulatory interactions at the distal ends of nucleosomes. They are found in a number of nucleosome-interacting proteins with very diverse architecture. CAP-Gly domains binding to core DNAOS bind to the motif sequence CAGAGG/EEFPA motif. In the case of alpha-tubulin, it is bound to a core of tetranucleotide sequence: LIG\_CAP-Gly\_1.

ELM(s):  
LIG\_CAP-Gly\_1 description:  
Short motifs are present, notably terminal sequence found in a small group of nucleosome-associated proteins. The EEFPA motif is highly conserved and can be divided into two distinct proteins, alpha-tubulin (EB proteins) and CLP170. The conserved sequence is mainly localized to the consensus sequence: [REDACTED], the unique region 'AF' is unique to alpha-tubulin. EB proteins and CLP170. These proteins are involved in various cellular processes such as chromatin organization, establishment and propagation of DNA, transcriptional regulation and cell division.

Pattern:  
[REDACTED]

Present in taxon(s):  
Eukaryota

Not represented in taxon(s):

See instances for LIG\_CAP-Gly\_1

Abstract:  
The nucleosome-associated protein glycinyl-rich domain (CAP-Gly) motif sequence motif (P117102) is a small, approximately 60-residue protein module essential in eukaryotes. CAP-Gly motifs have diverse functions in many proteins, including cytoplasmic inner proteins (CLP170 and CLP170L), the large subunit of the preinitiation complex (GCN5), the 300 kDa protein, localizing to centrosomes and centrosomes associated protein (300 kDa), the unique region 'AF' is unique to alpha-tubulin. EB proteins and CLP170. These proteins are involved in various cellular processes such as chromatin organization, establishment and propagation of DNA, transcriptional regulation and cell division.

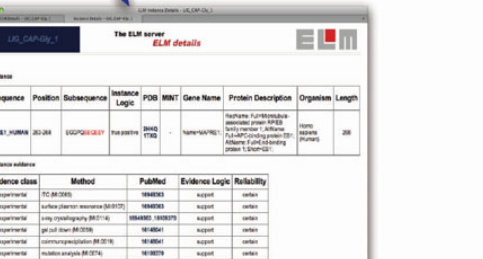
The CAP-Gly motif is a nucleosome-associated protein (NAP) motif found in the mammalian DNAOS motif which is responsible for targeting to the nucleosome EEFPA sequence motifs of CLP170, EB proteins, and alpha-tubulin. The CAP-Gly EEFPA function is essential for the recruitment of the preinitiation complex to CLP170 through 'AF' motif and for activation of CLP170. Furthermore, in mouse and human, the EEFPA motif of alpha-tubulin is subject to an enzymatic phosphorylation in various sites in the C-terminal region. This phosphorylation is subject to an enzymatic dephosphorylation in various sites in the C-terminal region. This phosphorylation is subject to an enzymatic dephosphorylation in various sites in the C-terminal region. This phosphorylation is subject to an enzymatic dephosphorylation in various sites in the C-terminal region.

Selected references:  
Cajigas R, et al. PNAS 113:14776-14781 (2016).  
Cajigas R, et al. PNAS 113:14776-14781 (2016).  
Cajigas R, et al. PNAS 113:14776-14781 (2016).

Scroll Down

Sequence	Position	Subsequence Click for evidence information	Instance Logic	PDB	Gene Name	Protein Description	Organism
TBA1A_HUMAN	613-621	2985644882EEF	True positive	2Z64	Nucleosome Factor 1A (Nucleosome Factor 1A) (Nucleosome Factor 1A)	Nucleosome Factor 1A (Nucleosome Factor 1A) (Nucleosome Factor 1A)	homo sapiens (Human)
CLP170_HUMAN	143-147	ATNCS02EF	True positive	3PFD	Nucleosome Factor 170 (Nucleosome Factor 170) (Nucleosome Factor 170)	Nucleosome Factor 170 (Nucleosome Factor 170) (Nucleosome Factor 170)	homo sapiens (Human)
MARE1_HUMAN	213-228	EQQG3EAEY	True positive	3M22	MARE1 (MARE1) (MARE1)	MARE1 (MARE1) (MARE1)	homo sapiens (Human)

Click on Link



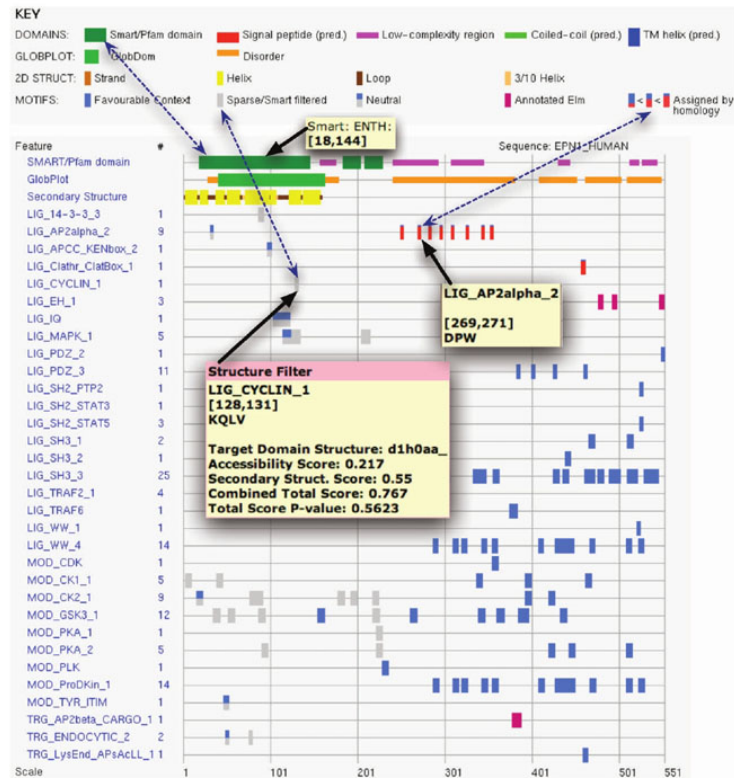
Instances

Sequence	Position	Subsequence	Instance Logic	PDB	MINT	Gene Name	Protein Description	Organism	Length
MARE1_HUMAN	213-228	EQQG3EAEY	True positive	3M22	1792	MARE1	MARE1 (MARE1) (MARE1)	homo sapiens (Human)	208

Evidence evidence

Evidence class	Method	PubMed	Evidence Logic	Reliability
experimental	DC 38 (2005)	1646363	support	certain
experimental	surface plasmon resonance (SPR)	1660293	support	certain
experimental	long range sequencing (LRS)	1660293; 1660297	support	certain
experimental	peptide array (PA)	1618541	support	certain
experimental	chromatin isolation (CI)	1618541	support	certain
experimental	nucleosome analysis (NA)	1618541	support	certain

**Figure 2.** Details from browse pages for the entry LIG\_CAP-Gly\_1 ([http://elm.eu.org/elmPages/LIG\\_CAP-Gly\\_1.html](http://elm.eu.org/elmPages/LIG_CAP-Gly_1.html)). The upper window shows the description and the regular expression for the motif. Scrolling down past the references and the GO terms (not shown) leads to the table of known instances (middle window). Key information in the table includes whether an instance is a true positive, a link to the UniProt sequence entry and, if available, links to PDB structure entries (49). Clicking on the linked sequence for the instance in the EB1 protein (MARE1\_HUMAN) opens a new page summarizing the annotated experimental evidence for the given instance. In this case, the motif has been exhaustively analysed and the supporting evidence is solid.



**Figure 3.** Graphic from the output page of the ELM server queried with Epsin-1 sequence from the UniProt entry EPN1\_HUMAN. The key indicates the content of the various coloured bars, e.g. the three connected by dotted arrows. Thirteen true LM instances are annotated either in this sequence or an orthologue from another species (magenta and red bar codes, respectively). Mouseover provides panels with different information depending on context, three examples of which are shown. One indicates an ENTH domain retrieved from SMART. A second points at an annotated DPW motif. The third mouseover provides the most detail: a structure for the ENTH domain (PDB entry d1h0) was used by the SF (41) to report that a cyclin motif candidate is too buried to be significant. Clicking on any object in the graphic will link to further details.

yeast, or cell lines from one of mouse, chicken or human. For a given LM class, the set of known instances may have been identified in a range of different species. Therefore, researchers are routinely faced with the issue of mapping experimental results from diverse organisms onto the protein sequence of their model organism. The instance mapper module addresses this issue for the ELM server.

A rarely used BLAST variant, PHI-BLAST, is at the core of the ELM instance mapper (64). PHI-BLAST

requires a regular expression in addition to the query sequence: the pattern must have at least one match in the query. We found PHI-BLAST to be ideally suited for mapping known LM matches from homologous sequences, so that the instance mapping issue was reduced to developing a protocol to utilize it effectively.

The flow scheme of the instance mapper is summarized in Figure 5. Sequences harbouring known instances are stored in a small BLAST formatted database. For each

Table 3. Web Service interfaces for the ELM tool suite

Resource module	Purpose of resource module	Links to WSDLs
ELM Database	Retrieve data stored by ELM	http://elm.eu.org/webservice/ELMdb.wsdl
ELMMatcher	Map ELM Motifs to query sequence	http://api.bioinformatics.org/wsdl/ELMdb.wsdl http://elm.eu.org/webservice/wsELMMatcher.wsdl http://api.bioinformatics.org/wsdl/ELMMatcher.wsdl
ELM CS Filter	Evaluate conservation of LM matches in reference sequence	http://conscore.embl.de/webservice/CS.wsdl
ELM SF	Evaluate accessibility and structure context of LM matches in query sequence given a reference structure	http://structurefilter.embl.de/webservice/structureFilter.wsdl
GlobPlot	Evaluate disorder propensity in query sequence	http://globplot.embl.de/webservice/globplot.wsdl
Phospho.ELM	Retrieve phosphorylation data stored by Phospho.ELM	http://phospho.elm.eu.org/webservice/phosphoELMdb.wsdl

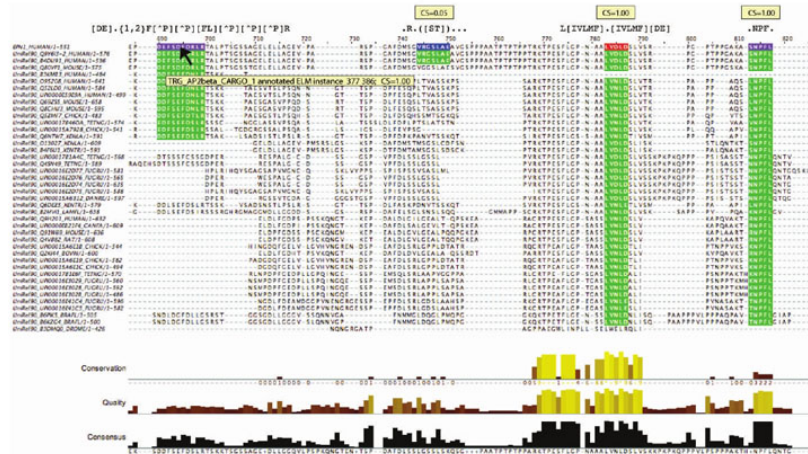


Figure 4. Representative results from the CS web interface, displayed with the annotated sequence alignment editor JalView (86). The alignment shows the set of sequences obtained by the CS filter with the human Epsin1 query sequence at top; the sequences belong to several paralogous families of Epsins. Four motif matches are highlighted in the reference sequence (magenta, annotated in this sequence; red, assigned by the instance mapper; blue, unannotated match) and in other sequences that align to the reference motif (green). The left-most match is a known instance of TRG\_AP2beta\_CARGO\_1 and gives a top score of 1.00 despite only being present in sequences belonging to two of the Epsin paralogs. This is because most sequences that lack the motif have gaps aligned to it that do not affect the CS score. The second motif is a candidate instance for MOD\_PKA\_2 but is poorly conserved, scoring 0.05. This candidate would probably not be worth investigating unless there was prior evidence for phosphorylation at the site. The remaining two motifs are known instances of LIG\_Clatrh\_ClatBox\_1 and LIG\_EH\_1, which obtain the maximum CS score since they are conserved in all Epsin paralogs.

pattern matching the query, this database is searched by PHI-BLAST. The instance mapper then parses the output and assigns a divergence-based score to any matches that are retrieved. These are then displayed in the ELM server graphical output (Figure 3).

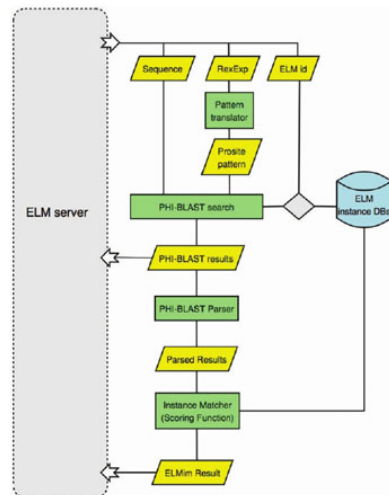
PHI-BLAST calculates an *E*-value, based on the BLAST bit score, which is useful for determining the statistical significance of a given alignment. However, this statistic does not reflect how similar the query sequence is to the LM instance sequence, which is particularly relevant for our purpose. To address this issue,

we have devised an ELM instance score  $S_{ei}$  that is calculated from the PHI-BLAST alignment:

$$S_{ei} = \frac{i - g/l_q}{\min(l_q, l_s)}$$

where *i* is the number of identical positions in the alignment, *g* is the number of gaps, *l<sub>q</sub>* is the length of the alignment (minus gaps), *l<sub>q</sub>* is the length of the query sequence and *l<sub>s</sub>* is the length of the subject sequence. The assumptions behind the score are that false matches are more likely at higher divergence and in longer





**Figure 5.** Flow scheme for the ELM Instance Mapper. For each predicted LM from an ELM database search, a PHI-BLAST search is performed against a database containing all sequences with known instances of the predicted LM. Input to PHI-BLAST is the query sequence and the ELM Regular Expression (which is adapted for use with PHI-BLAST). Each of the aligned motifs, between query and ELM instance sequence, are evaluated and scored (see main text). If the motif in the ELM instance sequence is a known instance, and the calculated score is above a threshold ( $S_{c} \geq 0.3$ ), it is reported as a mapped instance. Both the ELM instance mapper and the underlying PHI-BLAST results are returned to the ELM server, for the user to inspect.

sequences. At higher divergence, the sequences may be nonorthologous (or only partially so) or, in orthologous sequences, nonorthologous matches may also be superposed, especially for common, simple motifs. Therefore, while the instance matcher can retrieve genuine instances in sequences that are as low as 30% identity, a low score serves as a warning to evaluate the match. Note that this score is designed for evaluation of pairwise matches; if we had a multiple alignment and were confident that the alignment was correct for a motif, then the conservation can be scored as 'more' significant at higher divergence (61).

The instance mapper is a key addition to the resource as it unites the information content of the experimental instances stored in the ELM database with the motif exploration capabilities afforded by the ELM regular expressions.

#### USER COMMUNITY FEEDBACK AND INTERACTION

In common with other bioinformatics resources, only a few of the ELM users choose to communicate with us.

Users should know that certain types of communication are very useful to us. Obviously, if a server problem persists for a few hours, we should be informed immediately. Suggestions about the ELM resource interface would also be welcome—though we can probably only respond slowly to good ideas.

Of most use to ELM and the user community would be information to improve the data stored in ELM. Sometimes this might be a simple update such as an important instance that has been omitted, a new structure or a useful reference. More substantial help with creating or improving entries would be particularly valuable. In several cases, experts have contributed or reviewed entries for ELM. Entries with expert involvement include: LIG\_CAP-Gly\_1, LIG\_EH\_1, LIG\_SxIP\_EBH\_1, LIG\_ULM\_U2AF65\_1, LIG\_RRM\_PRI\_1, TRG\_AP2beta\_CARGO\_1 (65–70).

The obvious reason why researchers may be chary of getting involved with improving ELM is the time and effort that it costs. There is an upside that scientific information now disseminates to a great extent through the web: ELM can provide another route to showcase your work and, presumably, the prouder you are of your achievements, the more visible you would like them to be. We thank those researchers who have already helped us improve ELM and hope that their research will receive some reciprocal benefit.

#### ROLE OF ELM IN LM RESEARCH/DISCOVERY

As ELM has become more widely known to researchers, experimental investigations of candidate matches to known motifs have begun to appear in the literature. For example, an HCMV transmembrane protein has been shown to have LMs for cooption of cellular retention systems, aiding viral immune evasion (71). A candidate 14-3-3-binding phosphosite has been validated in the cytosolic C-terminus of integrin- $\alpha 4$  (72). Several regulatory motifs have been investigated in *Drosophila* cryptochrome, a regulator of circadian rhythm (73). Collectively such studies afford optimism that our work to establish the ELM resource will increasingly be justified by experimental application.

We take the view that by applying ELM ourselves, we can better evaluate and optimize our methodologies. We have sometimes been able to employ a protocol involving GO term enrichment to reveal sets of proteins with LM matches that are significantly enriched in specific contexts. Thus, we have reported a bioinformatics survey (63) of KEN box anaphase destruction motifs enriched in mitotic proteins: KEN box motifs in CHFR and C13orf3 are thought to aid in defining their roles in mitosis, though experimental validation is still needed (74,75). In a second example, while annotating the SUMO motif, we were able to define a larger motif, KEPE, superposed on a subset of sumoylation sites (62). It is, however, too soon for the role of KEPE to have been investigated.

The ELM instance dataset has been deployed by several bioinformatics groups in ways that have provided insight into LM context and/or to develop and benchmark

D176 *Nucleic Acids Research*, 2010, Vol. 38, Database issue

novel strategies for LM discovery. Thus, the anecdotal observation that LMs are more abundant in natively disordered protein sequence (21) has been verified by more systematic analyses using benchmarked native disorder predictors (28,29). More recently, this research line has been extended with the ANCHOR server providing benchmarked prediction of short stretches of sequence that have strong interacting potential (76). The local context of LMs has been further investigated, revealing that the adjacent peptide sequence often has a role in modulating LM function (77,78). Stemming from an awareness that viruses utilize numerous LMs to hijack cellular systems, Dinkel and Sticht (37) developed and benchmarked a pipeline to apply conservation and domain masking to motif candidates. Observing that multiple sequence alignment software has been over-trained on globular sequences and therefore performs quite poorly with short conserved motifs, the BALiBASE alignment benchmark suite was extended with an LM benchmark in the hope that this will lead to improved alignment algorithms (79).

While the ELM resource *per se* is not suited to *de novo* discovery of hitherto unknown motifs, the instances have been used by others to develop and benchmark tools for just this purpose. Yeast 2-hybrid data includes candidate LM-mediated interactions and both DILIMOT and SLiMFinder use interaction sets to search for enriched motifs in the binders of a protein (38,39,80). These methods depend on overrepresentation of a motif and therefore are probably not suited to motifs that have few biological instances. However, another promising approach uses amino acid preferences to sample 3D structural surfaces for sites with high peptide binding values (40): such methods have the potential to reveal LMs that have only a single functional instance in a proteome. These strategies illustrate how other data (interactions, structures) can be integrated into bioinformatics LM discovery pipelines, complementing experimental approaches for motif definition such as peptide libraries and arrays (81–83).

When we began the ELM project, LM bioinformatics was essentially nonexistent (21). The progress in the last few years has been impressive and exciting. There is growing awareness that the study of protein interactions is not just about globular-globular interfaces (5,84). Protein interaction data and domain surfaces can now be explored for possible LM interactors. There is much more to be done before researchers can pull up strong LM candidates as easily as running BLAST searches, but this goal—so important if we are to understand cell regulation—no longer seems to be impossibly fanciful.

#### EVALUATING AND APPLYING THE ELM SERVER RESULTS

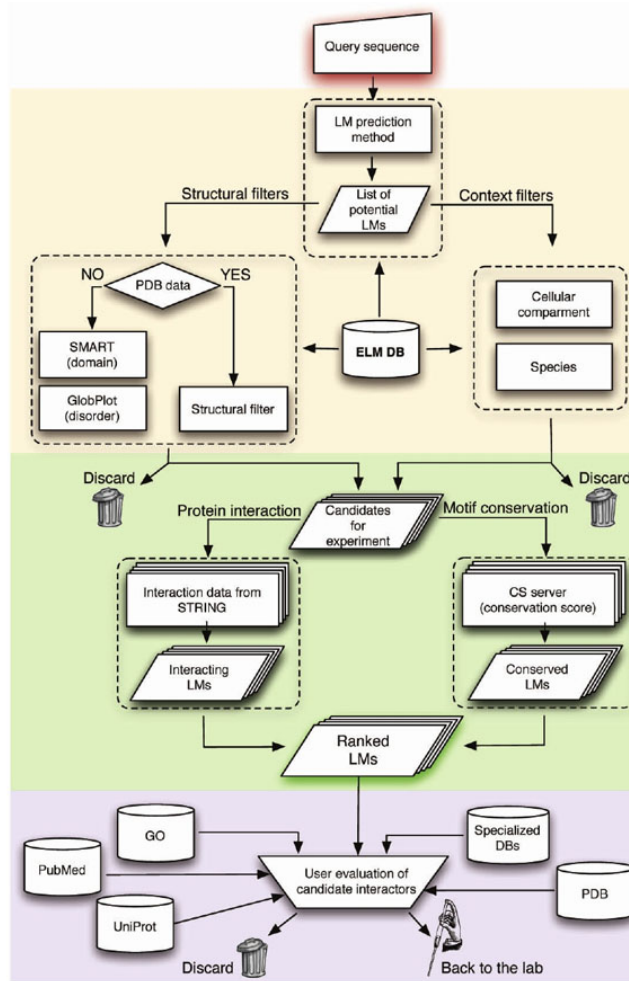
Candidate LMs require experimental validation. The key to using ELM is to select good candidates for experimental validation and not waste time on the poor ones. Since LMs are always interaction sites, they must be in the same

cell compartment as their ligand. There is little point in experimentally testing a candidate cyclin-binding motif in a collagen sequence. Likewise, a motif that is deeply buried in a solved structure makes a poor choice for experimentation (41). Therefore, it is first necessary to establish if a motif match is conserved, exposed and in the right cell compartment, according to the ELM filters. Motifs that pass these tests can then be further examined using a range of bioinformatics tools. Figure 6 shows a flowchart for how a typical motif evaluation might proceed. After the initial ELM tests, native disorder predictors and domain databases can give an indication of structural context. If the motif is within a known 3D structure, the context should be visualized; e.g. with PyMol (<http://pymol.sourceforge.net/>). Swiss-Prot features, the HPRD entry and phosphorylation databases may provide additional structure-function context. A user should always prepare a multiple sequence alignment and examine the motif conservation. Note that multiple alignment software sometimes struggle with motif alignments, with MAFFT (85) perhaps being the best current choice (79). If motifs are present but misaligned, an alignment editor such as JalView (86) may be helpful. Is the motif conserved in a specific lineage, e.g. vertebrates? If the motif is conserved, is the adjacent sequence less so? If things are looking good, it is important to ask whether the proposed LM function makes any sense for the protein; if this is unfamiliar, it is advisable to spend some time reading the literature: the ELM links to PubMed are a useful starting point, but unlikely to be exhaustive.

If LM candidates have survived the routine tests, there are other bioinformatics tools that might provide further insight. Protein interaction resources such as STRING (87), MINT (88) and IntAct (89) can reveal if a ligand protein is known to be close in the network. Interaction data can also be supplied to DILIMOT and/or SLiMFinder to evaluate whether there is statistical support for motif enrichment (38,39). Enrichment of motifs with UniProt GO terms and other keywords can sometimes provide statistical support for sets of motifs (62,63,90). SIRW is an online tool (<http://sirw.embl.de/index.html>) that allows keyword exploration for RegExps (91). If enrichment is found, SIRW can provide a probability estimate using Fisher's Exact Test. Of course, motif enrichment can be an artefact of sequence length or amino acid bias so judgement of the results is required. If the enriched set is not more conserved than the background, then it is unlikely to be biologically meaningful.

After doing all this, ask once again: Is the motif buried? We think it likely that inaccessible motifs are the most common reason for erroneous LM reports in the literature.

Even when an LM candidate is in the right cell compartment, and survives many other tests, it does not have to be functional as it still may never contact the ligand protein (20). There is increasing evidence that cell signalling decisions are made in large dynamic protein complexes. If a motif-containing protein is never in the same complex as a ligand protein, the motif will be false.



**Figure 6.** Workflow diagram illustrating how a user might explore LM candidates with ELM. The pipeline proceeds through three main phases utilizing the ELM resource (beige background) ELM associated tools (green) and more general bioinformatics resources (pink). Candidate LMs can be rejected by ELM filters if in unsuitable contexts. Sequence conservation and enrichment in interaction data using DiLiMot or SLiMFinder can provide additional scores to rank motifs. In the final phase any potentially relevant bioinformatics resources should be examined to provide further context to motif candidates. If promising candidates survive this process, the end point of the bioinformatics pipeline has been reached and laboratory validation is now required.

D178 *Nucleic Acids Research*, 2010, Vol. 38, Database issue

**Table 4.** The main experimental methods used in motif validation, as recorded in ELM

Experimental method	PSI-MI ID <sup>a</sup>	Number of occurrences
Mutation analysis	MI:0074	305
Pull down assay	MI:0096	200
Yeast 2 hybrid assay	MI:0018	115
Co-immunoprecipitation	MI:0019	98
X-ray crystallography	MI:0114	75
Motif Deletion	MI:0573	53
Competitive binding assay	MI:0405	39
Protein overlay assay	MI:0049	38
Colocalization by immunostaining	MI:0022	37
Nuclear magnetic resonance	MI:0077	30
Isothermal titration calorimetry (ITC)	MI:0065	29
Protein truncation mutants	MI:0422	28
Immunological detection and localization	MI:0427	27
Mass spectrometry	MI:0427	24
Motif transplantation		20
Western blot	MI:0113	19
Radiolabelling/pulse chase	MI:0517	19
Surface plasmon resonance	MI:0107	15

<sup>a</sup>Identifier for the HUPPO PSI-MI exchange standard entry that either defines or encompasses the listed experiment (92).

For this reason, cell localization assays are useful, although they can be misleading if overexpression is used. Coimmunoprecipitation and pull down experiments are also widely used as part of motif validation. We thought it might be of interest to list the most commonly annotated methods applied in motif validation and these are presented in Table 4. Since no one experiment is definitive, many of these methods will have been applied to a well-validated motif instance.

#### CURRENT LIMITATIONS AND FUTURE DIRECTIONS

In common with LM bioinformatics, in general, ELM has advanced to a state of practical usefulness, yet there is much more to do. LM RegExp matches cannot yet be taken as indicators of true functional sites and the candidates must be experimentally verified. The ELM dataset is incomplete with respect to motifs reported in the literature and there is work to be done to extend the coverage of the database: currently, users should not use ELM as a sole source of LM information. We have identified a need to improve the data captured regarding interactions of the ELM instances, which currently are of limited use for systems modelling *in silico*. ELM filtering can be improved in the short to medium term by embedding the CS filter and by using Swiss-Prot topology domains for automated cell compartment filtering of transmembrane proteins. In the ELM output, we would like to present the user with phosphorylation sites and other readily available information about the structure/function modules of query proteins. It is our hope that most of these goals will have been achieved when we next report on ELM.

#### ACKNOWLEDGEMENTS

The authors thank the former contributors to the ELM resource, the Bioinformatics developers who have applied the ELM instances to develop discovery methods and the ELM resource users whose web access statistics spurred us on.

#### FUNDING

The ELM Web Service interfaces were developed in the framework of the EU FP5 EMBRACE grant (LHSG-CT-2004-512092). The FIRB 2004 ITALBIONET grant (to A.V.); the NGFN DiGToP grant (to M.S.); the FP6 ProteomeBinders grant (to N.H.). SF development was aided by DAAD and Vigoni covered travel expenses between Heidelberg and Rome. Funding for open access charge: EMBL.

*Conflict of interest statement.* None declared.

#### REFERENCES

- Diella, F., Haslam, N., Chica, C., Budd, A., Michael, S., Brown, N.P., Trave, G. and Gibson, T.J. (2008) Understanding eukaryotic linear motifs and their role in cell signaling and regulation. *Front. Biosci.*, **13**, 6580–6603.
- Nedava, V. and Russell, R.B. (2006) Peptides mediating interaction networks: new leads at last. *Curr. Opin. Biotechnol.*, **17**, 465–471.
- Kadaveru, K., Vyas, J. and Schiller, M.R. (2008) Viral infection and human disease—insights from minimotifs. *Front. Biosci.*, **13**, 6455–6471.
- Fox-Erlich, S., Schiller, M.R. and Gryk, M.R. (2009) Structural conservation of a short, functional, peptide-sequence motif. *Front. Biosci.*, **14**, 1143–1151.
- Petsalaki, E. and Russell, R.B. (2008) Peptide-mediated interactions in biological systems: new discoveries and applications. *Curr. Opin. Biotechnol.*, **19**, 344–350.
- Chen, Y., Yang, Y., van Overbeek, M., Donigian, J.R., Baciu, P., de Lange, T. and Lei, M. (2008) A shared docking motif in TRF1 and TRF2 used for differential recruitment of telomeric proteins. *Science*, **319**, 1092–1096.
- Salsmann, A., Schaffner-Reckinger, E. and Kieffer, N. (2006) RGD, the Rho'd to cell spreading. *Eur. J. Cell Biol.*, **85**, 249–254.
- Pawson, T. and Nash, P. (2003) Assembly of cell regulatory systems through protein interaction domains. *Science*, **300**, 445–452.
- Hilser, V.J. and Thompson, E.B. (2007) Intrinsic disorder as a mechanism to optimize allosteric coupling in proteins. *Proc. Natl Acad. Sci. USA*, **104**, 8311–8315.
- Wright, P.E. and Dyson, H.J. (2009) Linking folding and binding. *Curr. Opin. Struct. Biol.*, **19**, 31–38.
- Mayer, B.J., Blinov, M.L. and Loew, L.M. (2009) Molecular machines or pleomorphic ensembles: signaling complexes revisited. *J. Biol.*, **8**, 81.
- Stein, A., Pache, R.A., Bernado, P., Pons, M. and Aloy, P. (2009) Dynamic interactions of proteins in complex networks: a more structured view. *FEBS J.*, **276**, 5390–5405.
- Kitano, H. (2007) Towards a theory of biological robustness. *Mol. Syst. Biol.*, **3**, 137.
- Pawson, T. and Kofler, M. (2009) Kinome signaling through regulated protein-protein interactions in normal and cancer cells. *Curr. Opin. Cell Biol.*, **21**, 147–153.
- Smock, R.G. and Gierasch, L.M. (2009) Sending signals dynamically. *Science*, **324**, 198–203.
- Volonte, C., D'Ambrosi, N. and Amadio, S. (2008) Protein cooperation: from neurons to networks. *Prog. Neurobiol.*, **86**, 61–71.
- Whitty, A. (2008) Cooperativity and biological complexity. *Nat. Chem. Biol.*, **4**, 435–439.



18. Williamson, J.R. (2008) Cooperativity in macromolecular assembly. *Nat. Chem. Biol.*, **4**, 458–465.
19. Tan, C.S., Bodenmiller, B., Pasculescu, A., Jovanovic, M., Hengartner, M.O., Jorgensen, C., Bader, G.D., Aebersold, R., Pawson, T. and Lindling, R. (2009) Comparative analysis reveals conserved protein phosphorylation networks implicated in multiple diseases. *Sci. Signal.*, **2**, ra39.
20. Gibson, T.J. (2009) Cell regulation: determined to signal discrete cooperation. *Trends Biochem. Sci.*, **34**, 471–482.
21. Puntervoll, P., Lindling, R., Gemünd, C., Chabanis-Davidson, S., Mattingsdal, M., Cameron, S., Martin, D.M., Ausiello, G., Brannetti, B., Costantini, A. et al. (2003) ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Res.*, **31**, 3625–3630.
22. Rajasekaran, S., Balla, S., Gradie, P., Gryk, M.R., Kadaveru, K., Kundeti, V., Maciejewski, M.W., Mi, T., Rubino, N., Vyas, J. et al. (2009) Minimotif mimer 2nd release: a database and web system for motif search. *Nucleic Acids Res.*, **37**, D185–D190.
23. Hornbeck, P.V., Chhabra, I., Kornhauser, J.M., Skrzypek, E. and Zhang, B. (2004) PhosphoSite: A bioinformatics resource dedicated to physiological protein phosphorylation. *Proteomics*, **4**, 1551–1561.
24. Diella, F., Gould, C.M., Chica, C., Via, A. and Gibson, T.J. (2008) Phospho.ELM: a database of phosphorylation sites—update 2008. *Nucleic Acids Res.*, **36**, D240–D244.
25. Gnad, F., Ren, S., Cox, J., Olsen, J.V., Macek, B., Orosi, M. and Mann, M. (2007) PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. *Genome Biol.*, **8**, R250.
26. Keshava Prasad, T.S., Goel, R., Kandhasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A. et al. (2009) Human Protein Reference Database—2009 update. *Nucleic Acids Res.*, **37**, D767–D772.
27. UniProt Consortium. (2009) The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res.*, **37**, D169–D174.
28. Fuxreiter, M., Tompa, P. and Simon, I. (2007) Local structural disorder imparts plasticity on linear motifs. *Bioinformatics*, **23**, 950–956.
29. Ren, S., Uversky, V.N., Chen, Z., Dunker, A.K. and Obradovic, Z. (2008) Short Linear Motifs recognized by SH2, SH3 and Ser/Thr Kinase domains are conserved in disordered protein regions. *BMC Genomics*, **9**(Suppl. 2), S26.
30. Russell, R.B. and Gibson, T.J. (2008) A careful disorderliness in the proteome: sites for interaction and targets for future therapies. *FEBS Lett.*, **582**, 1271–1275.
31. Bourhis, J.M., Canard, B. and Longhi, S. (2007) Predicting protein disorder and induced folding: from theoretical principles to practical applications. *Curr. Protein Pept. Sci.*, **8**, 135–149.
32. He, B., Wang, K., Liu, Y., Xue, B., Uversky, V.N. and Dunker, A.K. (2009) Predicting intrinsic disorder in proteins: an overview. *Cell Res.*, **19**, 929–949.
33. Finn, R.D., Tate, J., Mistry, J., Coghill, P.C., Sammut, S.J., Hotz, H.R., Ceric, G., Forslund, K., Eddy, S.R., Sonnhammer, E.L. et al. (2008) The Pfam protein families database. *Nucleic Acids Res.*, **36**, D281–D288.
34. Letunic, I., Doerks, T. and Bork, P. (2009) SMART 6: recent updates and new developments. *Nucleic Acids Res.*, **37**, D229–D232.
35. Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., Cuče, B.A., de Castro, E., Lachaize, C., Langendijk-Genevaux, P.S. and Sigrist, C.J. (2008) The 20 years of PROSITE. *Nucleic Acids Res.*, **36**, D245–D249.
36. Hunter, S., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L. et al. (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res.*, **37**, D211–D215.
37. Dinkel, H. and Stich, H. (2007) A computational strategy for the prediction of functional linear peptide motifs in proteins. *Bioinformatics*, **23**, 3297–3303.
38. Edwards, R.J., Davey, N.E. and Shields, D.C. (2007) SLIMFinder: a probabilistic method for identifying over-represented, convergently evolved, short linear motifs in proteins. *PLoS ONE*, **2**, e967.
39. Neduva, V. and Russell, R.B. (2006) DILIMOT: discovery of linear motifs in proteins. *Nucleic Acids Res.*, **34**, W350–W355.
40. Petsalaki, E., Stark, A., Garcia-Urdiales, E. and Russell, R.B. (2009) Accurate prediction of peptide binding sites on protein surfaces. *PLoS Comput. Biol.*, **5**, e1000335.
41. Via, A., Gould, C.M., Gemünd, C., Gibson, T.J. and Helmer-Citterich, M. (2009) A structure filter for the Eukaryotic Linear Motif Resource. *BMC Bioinformatics*, **10**, 351.
42. Hunt, T. (1990) Protein sequence motifs involved in recognition and targeting: a new series. *Trends Biochem. Sci.*, **15**, 305.
43. Pelham, H.R. (1990) The retention signal for soluble proteins of the endoplasmic reticulum. *Trends Biochem. Sci.*, **15**, 483–486.
44. Dingwall, C. and Laskey, R.A. (1991) Nuclear targeting sequences – a consensus? *Trends Biochem. Sci.*, **16**, 478–481.
45. Glotzer, M., Murray, A.W. and Kirschner, M.W. (1991) Cyclin is degraded by the ubiquitin pathway. *Nature*, **349**, 132–138.
46. Dice, J.F. (1990) Peptide sequences that target cytosolic proteins for lysosomal proteolysis. *Trends Biochem. Sci.*, **15**, 305–309.
47. Hantschel, O., Nagar, B., Guettler, S., Kretschmar, J., Dorey, K., Kurryan, J. and Superti-Furga, G. (2003) A myristoyl phosphorytyrosine switch regulates c-Abl. *Cell*, **112**, 845–857.
48. Kadlec, J., Izaurralde, E. and Cusack, S. (2004) The structural basis for the interaction between nonsense-mediated mRNA decay factors UPF2 and UPF3. *Nat. Struct. Mol. Biol.*, **11**, 330–337.
49. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
50. Gene Ontology Consortium. (2008) The Gene Ontology project in 2008. *Nucleic Acids Res.*, **36**, D440–D444.
51. Sayers, E.W., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Chetverin, V., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S. et al. (2009) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **37**, D5–15.
52. Steinmetz, M.O. and Akhmanova, A. (2008) Capturing protein tails by CAP-Gly domains. *Trends Biochem. Sci.*, **33**, 535–545.
53. Chenna, R. and Gemünd, C. (2000) egimodel: CGI programming made easy with Python. *Linux J.*, **75**, 142–149.
54. Eddy, S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
55. Krogh, A. (2008) What are artificial neural networks? *Nat. Biotechnol.*, **26**, 195–197.
56. Seiler, M., Mehrle, A., Poustka, A. and Wiemann, S. (2006) The 3of5 web application for complex and comprehensive pattern matching in protein sequences. *BMC Bioinformatics*, **7**, 144.
57. Obenaus, J.C., Cantley, L.C. and Yaffe, M.B. (2003) ScanSite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res.*, **31**, 3635–3641.
58. Miller, M.L., Jensen, L.J., Diella, F., Jorgensen, C., Tinti, M., Li, L., Hsiung, M., Parker, S.A., Bordeaux, J., Sichert-Ponten, T. et al. (2008) Linear motif atlas for phosphorylation-dependent signaling. *Sci. Signal.*, **1**, ra2.
59. Pettifer, S., Thorne, D., McDermott, P., Attwood, T., Baran, J., Bryne, J.C., Hupponen, T., Mowbray, D. and Vriend, G. (2009) An active registry for bioinformatics web services. *Bioinformatics*, **25**, 2090–2091.
60. Snedley, D., Haider, S., Ballester, B., Holland, R., London, D., Thorisson, G. and Kasprzyk, A. (2009) BioMart—biological queries made easy. *BMC Genomics*, **10**, 22.
61. Chica, C., Labarga, A., Gould, C.M., Lopez, R. and Gibson, T.J. (2008) A tree-based conservation scoring method for short linear motifs in multiple alignments of protein sequences. *BMC Bioinformatics*, **9**, 229.
62. Diella, F., Chabanis, S., Luck, K., Chica, C., Ramu, C., Nerlov, C. and Gibson, T.J. (2009) KEPE—a motif frequently superimposed on sumoylation sites in metazoan chromatin proteins and transcription factors. *Bioinformatics*, **25**, 1–5.
63. Michael, S., Trave, G., Ramu, C., Chica, C. and Gibson, T.J. (2008) Discovery of candidate KEN-box motifs using cell cycle keyword enrichment combined with native disorder prediction and motif conservation. *Bioinformatics*, **24**, 453–457.
64. Zhang, Z., Schaffer, A.A., Miller, W., Madden, T.L., Lipman, D.J., Koonin, E.V. and Altschul, S.F. (1998) Protein sequence similarity searches using patterns as seeds. *Nucleic Acids Res.*, **26**, 3986–3990.
65. Weisbrich, A., Honnappa, S., Jaussi, R., Okhrimenko, O., Frey, D., Jelezarov, I., Akhmanova, A. and Steinmetz, M.O. (2007)

D180 *Nucleic Acids Research*, 2010, Vol. 38, Database issue

- Structure-function relationship of CAP-Gly domains. *Nat. Struct. Mol. Biol.*, **14**, 959–967.
66. Rumpf, J., Simon, B., Jung, N., Maritzen, T., Hauke, V., Sattler, M. and Groemping, Y. (2008) Structure of the Eps15-stonin2 complex provides a molecular explanation for EH-domain ligand specificity. *EMBO J.*, **27**, 558–569.
  67. Honnappa, S., Gouveia, S.M., Weisbrich, A., Damberger, F.F., Bhavesh, N.S., Jawhari, H., Grigoriev, I., van Rijssel, F.J., Buey, R.M., Lawera, A. *et al.* (2009) An EB1-binding motif acts as a microtubule tip localization signal. *Cell*, **138**, 366–376.
  68. Corsini, L., Bonnal, S., Basquin, J., Hothorn, M., Scheffzek, K., Valcarcel, J. and Sattler, M. (2007) U2AF-homology motif interactions are required for alternative splicing regulation by SPF45. *Nat. Struct. Mol. Biol.*, **14**, 620–629.
  69. Rideau, A.P., Gooding, C., Simpson, P.J., Monic, T.P., Lorenz, M., Huttelmaier, S., Singer, R.H., Matthews, S., Curry, S. and Smith, C.W. (2006) A peptide motif in Raver1 mediates splicing repression by interaction with the PTB RRM2 domain. *Nat. Struct. Mol. Biol.*, **13**, 839–848.
  70. Edeling, M.A., Mishra, S.K., Keyel, P.A., Steinhauser, A.L., Collins, B.M., Roth, R., Heuser, J.E., Owen, D.J. and Traub, L.M. (2006) Molecular switches involving the AP-2 beta2 appendage regulate endocytic cargo selection and clathrin coat assembly. *Dev. Cell*, **10**, 329–342.
  71. Maffei, M., Ghiotto, F., Occhino, M., Bono, M., De Santanna, A., Battini, L., Gusella, G.L., Fais, F., Bruno, S. and Ciccone, E. (2008) Human cytomegalovirus regulates surface expression of the viral protein UL18 by means of two motifs present in the cytoplasmic tail. *J. Immunol.*, **180**, 969–979.
  72. Deakin, N.O., Bass, M.D., Warwood, S., Schoelermann, J., Mostafavi-Pour, Z., Knight, D., Ballestrin, C. and Humphries, M.J. (2009) An integrin- $\alpha$ 4-14-3-3(zeta)-paxillin ternary complex mediates localised Cdc42 activity and accelerates cell migration. *J. Cell Sci.*, **122**, 1654–1664.
  73. Hemsley, M.J., Mazzotta, G.M., Mason, M., Dissel, S., Toppo, S., Pagano, M.A., Sandrelli, F., Meggio, F., Rosato, E., Costa, R. *et al.* (2007) Linear motifs in the C-terminus of D. melanogaster cryptochrome. *Biochem. Biophys. Res. Commun.*, **355**, 531–537.
  74. Privette, L.M., Weier, J.F., Nguyen, H.N., Yu, X. and Petty, E.M. (2008) Loss of CHFR in human mammary epithelial cells causes genomic instability by disrupting the mitotic spindle assembly checkpoint. *Neoplasia*, **10**, 643–652.
  75. Theis, M., Slabicki, M., Junqueira, M., Paszkowski-Rogacz, M., Sontheimer, J., Kittler, R., Heninger, A.K., Glatter, T., Krausma, K., Poser, I. *et al.* (2009) Comparative profiling identifies Cl3orf3 as a component of the Ska complex required for mammalian cell division. *EMBO J.*, **28**, 1453–1465.
  76. Meszaros, B., Simon, I. and Dosztanyi, Z. (2009) Prediction of protein binding regions in disordered proteins. *PLoS Comput. Biol.*, **5**, e1000376.
  77. Stein, A. and Aloy, P. (2008) Contextual specificity in peptide-mediated protein interactions. *PLoS ONE*, **3**, e2524.
  78. Chica, C., Diella, F. and Gibson, T.J. (2009) Evidence for the concerted evolution between short linear protein motifs and their flanking regions. *PLoS ONE*, **4**, e6052.
  79. Perrodou, E., Chica, C., Poch, O., Gibson, T.J. and Thompson, J.D. (2008) A new protein linear motif benchmark for multiple sequence alignment software. *BMC Bioinformatics*, **9**, 213.
  80. Neduva, V., Linding, R., Su-Angrand, I., Stark, A., de Masi, F., Gibson, T.J., Lewis, J., Serrano, L. and Russell, R.B. (2005) Systematic discovery of new recognition peptides mediating protein interaction networks. *PLoS Biol.*, **3**, e405.
  81. Ferraro, E., Via, A., Ausiello, G. and Helmer-Citterich, M. (2005) A neural strategy for the inference of SH3 domain-peptide interaction specificity. *BMC Bioinformatics*, **6**(Suppl. 4), S13.
  82. Machida, K., Thompson, C.M., Dierck, K., Jablonowski, K., Karkkainen, S., Liu, B., Zhang, H., Nash, P.D., Newman, D.K., Nollau, P. *et al.* (2007) High-throughput phosphotyrosine profiling using SH2 domains. *Mol. Cell*, **26**, 899–915.
  83. Zhu, G., Fujii, K., Liu, Y., Codrea, V., Herrero, J. and Shaw, S. (2005) A single pair of acidic residues in the kinase major groove mediates strong substrate preference for P-2 or P-5 arginine in the AGC, CAMK, and STE kinase families. *J. Biol. Chem.*, **280**, 36372–36379.
  84. Stein, A., Panjkovich, A. and Aloy, P. (2009) 3did Update: domain-domain and peptide-mediated interactions of known 3D structure. *Nucleic Acids Res.*, **37**, D300–D304.
  85. Katoh, K. and Toh, H. (2008) Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform.*, **9**, 286–298.
  86. Waterhouse, A.M., Procter, J.B., Martin, D.M., Clamp, M. and Barton, G.J. (2009) Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**, 1189–1191.
  87. Jensen, L.J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., Doerks, T., Julien, P., Roth, A., Simonovic, M. *et al.* (2009) STRING 8 – a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.*, **37**, D412–D416.
  88. Chatri-aryamontri, A., Ceol, A., Palazzi, L.M., Nardelli, G., Schneider, M.V., Castagnoli, L. and Cesareni, G. (2007) MINT: the Molecular Interaction database. *Nucleic Acids Res.*, **35**, D572–D574.
  89. Kerrien, S., Alam-Faruque, Y., Aranda, B., Bancarz, I., Bridge, A., Derow, C., Dimmer, E., Feuerhahn, M., Friedrichsen, A., Huntley, R. *et al.* (2007) IntAct—open source resource for molecular interaction data. *Nucleic Acids Res.*, **35**, D561–D565.
  90. Copley, R.R. (2005) The EH1 motif in metazoan transcription factors. *BMC Genomics*, **6**, 169.
  91. Ramu, C. (2003) SIRW: A web server for the Simple Indexing and Retrieval System that combines sequence motif searches with keyword searches. *Nucleic Acids Res.*, **31**, 3771–3774.
  92. Hermjakob, H., Montecchi-Palazzi, L., Bader, G., Wojcik, J., Salwinski, L., Ceol, A., Moore, S., Orchard, S., Sarkans, U., von Mering, C. *et al.* (2004) The HUPPO PSI's molecular interaction format—a community standard for the representation of protein interaction data. *Nat. Biotechnol.*, **22**, 177–183.

## **Paper II: Exploiting Biological and Biochemical Information**

## Exploiting Publicly Available Biological and Biochemical Information for the Discovery of Novel Short Linear Motifs

Ahmed Sayadi<sup>1</sup>, Leonardo Briganti<sup>2</sup>, Anna Tramontano<sup>1,3</sup>, Allegra Via<sup>1\*</sup>

<sup>1</sup> Department of Physics, Sapienza University of Rome, Rome, Italy, <sup>2</sup> Department of Biology, University of Rome "Tor Vergata", Rome, Italy, <sup>3</sup> Istituto Pasteur Fondazione Cenci Bolognetti, Sapienza University of Rome, Rome, Italy

### Abstract

The function of proteins is often mediated by short linear segments of their amino acid sequence, called Short Linear Motifs or SLiMs, the identification of which can provide important information about a protein function. However, the short length of the motifs and their variable degree of conservation makes their identification hard since it is difficult to correctly estimate the statistical significance of their occurrence. Consequently, only a small fraction of them have been discovered so far. We describe here an approach for the discovery of SLiMs based on their occurrence in evolutionarily unrelated proteins belonging to the same biological, signalling or metabolic pathway and give specific examples of its effectiveness in both rediscovering known motifs and in discovering novel ones. An automatic implementation of the procedure, available for download, allows significant motifs to be identified, automatically annotated with functional, evolutionary and structural information and organized in a database that can be inspected and queried. An instance of the database populated with pre-computed data on seven organisms is accessible through a publicly available server and we believe it constitutes by itself a useful resource for the life sciences (<http://www.biocomputing.it/modipath>).

**Citation:** Sayadi A, Briganti L, Tramontano A, Via A (2011) Exploiting Publicly Available Biological and Biochemical Information for the Discovery of Novel Short Linear Motifs. PLoS ONE 6(7): e22270. doi:10.1371/journal.pone.0022270

**Editor:** Christos A. Ouzounis, The Centre for Research and Technology, Hellas, Greece

**Received:** February 3, 2011; **Accepted:** June 22, 2011; **Published:** July 20, 2011

**Copyright:** © 2011 Sayadi et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Funding:** This work was partially supported by Award No. KUK-I1-012-43 made by King Abdullah University of Science and Technology (KAUST: <http://www.kaust.edu.sa/>), by Fondazione Roma (<http://www.fondazioneroma.it/it/index.html>) and by the Italian Ministry of Health (<http://www.salute.gov.it/>), contract no. onc\_ord 25/07, FIRB ITALBIONET and PROTEOMICA. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

\* E-mail: [allegra.via@uniroma1.it](mailto:allegra.via@uniroma1.it)

### Introduction

Short Linear Motifs (SLiMs) are sub-sequences of few adjacent amino acids (typically between three and ten residues in length) contributing to the molecular function of proteins. SLiMs have been estimated to mediate 15%–40% of protein-protein interactions [1,2] and recognized to be critical for many biological processes (e.g. sub-cellular targeting, post-translational modification, signal transduction, etc.) [3]. Protein domain-SLiM interactions have also been linked to several diseases, such as Alzheimer [4] and Huntington [5] diseases, Muscular Dystrophy [6], and malaria [7,8]. Examples of SLiMs are the C-Mannosylation site WxxW [9], the PxxP SH3 domain binding motif [10,11], the KDELF Golgi-to-Endoplasmic Reticulum retrieving signal [12], the polyproline rich peptides interacting with WW domains [13] and phosphorylation sites [14]. Given their short length, their variable degree of conservation (positions may be degenerate in terms of permitted amino acids), their weak binding affinity [1], the difficulty of correctly estimate the statistical significance of their occurrence in protein sequences, and the fact that most of them reside in disordered regions [15], SLiMs are difficult to discover both experimentally and computationally (e.g. [16]). For this reason, only few hundreds of motifs are known as of today while it is believed that the majority of SLiMs have still to be discovered (e.g. [16]).

Most of the known SLiMs are deposited in manually annotated repositories including PROSITE [17], ELM [18] and MnM [19]. The manual annotation of motifs is an important process that, besides being instrumental as a guide to experimentalists, allows the construction of benchmarking datasets necessary for the assessment of the performance of motif prediction tools. It is however difficult if not impossible to scale the manual process at the level required for handling high throughput data. The cogent need for *de novo* discovery of SLiMs has prompted the development of automatic motif discovery approaches that can be broadly divided into two types: those that use sequence alignments to identify motifs in evolutionarily related proteins (e.g. MEME [20]) and those that use over-representation of motifs in evolutionarily unrelated proteins sharing a common functional characteristic. For example, DILL-MOT [21], which is based on the TEIRESIAS [22] combinatorial pattern discovery algorithm, searches over-represented motifs in non-homologous proteins with a common interaction partner. The MoVIN server [23] is based on the same principle and identifies the presence of common motifs in proteins interacting with the same partner. SLiMDisc [24] uses TEIRESIAS to find shared motifs in all (homologous and non-homologous) proteins with a common attribute (biological function, sub-cellular location, or a common interaction partner); identified common substrings are subsequently weighted according to the evolutionary relationships of the proteins containing the motif.



SLiMFinder [25] is a combined software package that implements two algorithms, SLiMBuilder and SLiMChance. The former is designed to identify motifs that are shared by unrelated proteins whereas the latter calculates a score that accounts for the probability that a given motif occurs in a dataset of unrelated proteins by chance. In practice, the motifs identified by SLiMBuilder are returned with a significance value provided by SLiMChance. SLiMFinder allows the search to be restricted to specific regions of the set of input proteins such as disordered or non-disordered subsequences, positions annotated by UniProt features and low complexity regions.

The rationale behind most available SLiM discovery systems is the assumption that motifs mediate transient interactions, and therefore play a key role in signalling pathways, the proteins of which often contain (e.g.) SH2, SH3, PTB, 14-3-3 domain interacting motifs. Less well established is whether SLiMs are equally important in mediating interactions in metabolic pathways, which is in principle very likely. In a metabolic pathway a principal chemical is modified by a series of reactions carried out by the proteins of the pathway which therefore interact with either the principal chemical or one of its derivatives. Furthermore, specific reactions in a metabolic pathway are temporally and spatially compartmentalized [26].

It is therefore reasonable to expect that the corresponding proteins and enzymes, or a subset of them, may share a binding motif and/or one or more common cellular localization motifs and that the inspection of the sequence of proteins involved in a common pathway might be very useful for the discovery of novel functional motifs. This is the strategy followed by the procedure described here and we show that it is indeed possible to discover novel motifs shared by proteins involved in the same biological (signalling or metabolic) pathway.

In our procedure, named MoDiPath, proteins are grouped according to the KEGG Pathway Database [27]. The database contains both metabolic pathways (e.g. fatty acid biosynthesis, purine metabolism), based on indirect protein-protein interactions, and non-metabolic pathways (e.g. secretory, signaling pathways), based on direct protein-protein interactions.

MoDiPath identifies over-represented SLiMs in KEGG pathways in different organisms, and uses functional and structural annotation to assess their plausibility. By applying this protocol to seven organisms, we could both re-discover previously known motifs and detect several novel ones. The discovered motifs, annotated with functional, structural and evolutionary conservation information and linked to several other SLiM resources, are stored in a publicly available database accessible through a Web interface (<http://www.biocomputing.it/modipath>).

The automatic procedure can be downloaded from <http://www.biocomputing.it/modipath/MoDiPath.11-04-2011.zip> and installed locally.

## Results

### The MoDiPath procedure

The MoDiPath procedure is designed to search for motifs that are over-represented in a set of unrelated proteins belonging to the same biological pathway.

We applied the procedure to all KEGG pathways from seven organisms (*H.sapiens*, *R.norvegicus*, *M.musculus*, *D.melanogaster*, *C.elegans*, *S.cerevisiae*, *E.coli*) and made these pre-computed data available via a web server.

The pipeline consists of the following steps (see Figure 1 and Materials and Methods):

- 1) Sets of proteins belonging to a given pathway in a given organism are collected (Table S1.1 and S1.2);
- 2) The proteins are filtered to restrict the analysis to proteins that share not more than 40% and 25% sequence identity and that are therefore less likely to be evolutionarily related (Table S1.1 and S1.2). The 25% threshold was selected since it is commonly used for safe removal of homologous proteins (e.g. [28,29]). We also allow the user to increase the threshold up to 40%, the lower level of redundancy used, for example, by CD-HIT [30];
- 3) the SLiMFinder algorithm is used for the identification of over-represented SLiMs shared by all (or a subset of) non-redundant proteins belonging to the pathway;
- 4) the specificity of the identified motifs is assessed by comparing the number of motif occurrences in the set of proteins belonging to the pathway with that obtained from searching the motifs in the whole set of KEGG protein sequences and in the UniProt knowledge database [31];
- 5) motifs are ranked based on their hyper-geometric p-value (see later) and pathway-specific ones are identified;
- 6) motifs are compared with known SLiMs in other databases and annotated with functional, structural and evolutionary conservation information;
- 7) the annotated motifs are stored in the MoDiPath database.

### Re-discovered and newly discovered motifs

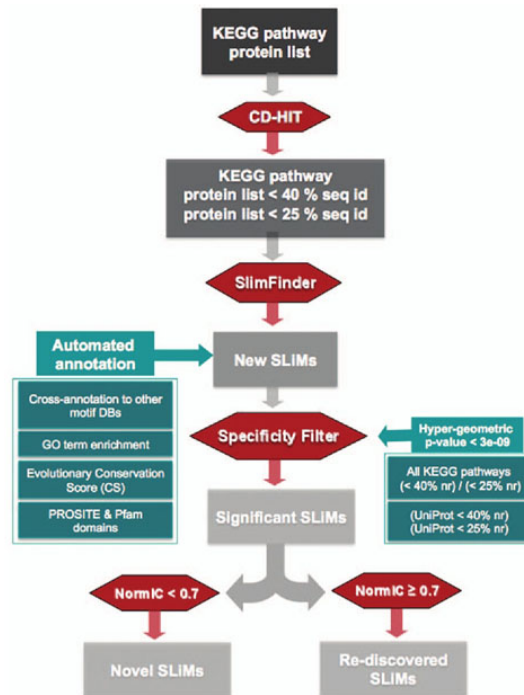
The MoDiPath procedure was able to uncover and re-discover a significant number of motifs (Table 1).

We found 104 statistically significant motifs specific to human pathways (21 in metabolic and 83 in non-metabolic pathways). Out of these 104 motifs, 82 have some degree of similarity to already known motifs present in other databases. We define two motifs to be similar if their CompariMotif score [32] is above 0.7 (see Materials and Methods). CompariMotif takes into account exact matches, variants of degenerate motifs and complex overlapping motifs.

Sixty-three of these motifs are identical to known motifs stored in one of the following databases: ELM [18], MnM [19], PhosphoMotif Finder [33], a dataset of motifs extracted from the literature, and a set of SLiMs predicted by Neduva and Russell [1]. Interestingly, twenty-two SLiMs are novel and share no similarity with any known motif. Table 1 shows the number of detected SLiMs already present in existing databases or very similar to one of their entries as well as the number of newly discovered SLiMs in each analysed organism. Novel motifs are reported in Table S2.1 (novel motifs detected in the 25% non-redundant dataset of sequences) and Table S2.2 (novel motifs detected in the 40% non-redundant dataset of sequences) and re-discovered motifs are reported in Table S3.1 (known motifs detected in the 25% non-redundant dataset of sequences) and Table S3.2 (known motifs detected in the 40% non-redundant dataset of sequences).

Table 2 reports the total number of KEGG pathways analysed per species and the number of pathways for which at least one SLiM has been detected.

Motifs were also compared to each other (all-against-all) in order to group similar motifs identified by CompariMotif (CompariMotif score  $\geq 0.7$ ). The data reported in Table 1 were filtered by taking into account only one representative motif (motif *representative*) for each similarity group and the results are shown in Table 3, from which it can be appreciated that there are 64 statistically significant motifs specific for human pathways (18 in



**Figure 1. Flowchart of the MoDiPath procedure.** NormIC is the CompariMotif [32] similarity score. The CompariMotif tool was used to find similarities between motifs automatically discovered by MoDiPath and motifs already annotated in other databases. doi:10.1371/journal.pone.0022270.g001

metabolic and 46 in non-metabolic pathways). More detailed information obtained from the all-against-all motif comparison is reported in Tables S2.1 and S2.2 (for novel motifs) and Tables S3.1 and S3.2 (for known motifs).

Data reported in Tables 1, 2, and 3 refer to the 40% non-redundant sequence dataset. The corresponding data for the 25% non-redundant dataset can be found in the Supporting Information S1 file.

#### Evolutionary conservation of SLiMs

Evolutionary conservation is often used for assessing the biological significance of predicted SLiMs. It is reasonable to expect that if the residues composing a motif have a functional role, the motif is evolutionary conserved. On the other hand, SLiMs are usually short, tend to localise in disordered regions that are difficult to align, and might not be shared even by closely related sequences as a result of single mutations. These observations imply that it is difficult to trace their evolutionary history. Here, we use a scoring

scheme that has been specifically designed for SLiMs [34] taking into account the potential problems mentioned above. We also used the CompariMotif algorithm to highlight motifs that are shared by two or more of the species under study (Table S4.1 and Table S4.2). We found that, with some exceptions, motifs shared by different organisms are related to similar or identical pathways. Fifty-five (45 known and 10 novel) out of the 104 human specific motifs are shared by proteins belonging to the same pathway in at least another species in the 40% sequence dataset (Table S4.2).

#### Assessment of some re-discovered and newly discovered motifs

We manually analysed a number of examples extracted from the list of re-discovered SLiMs (Table S3.1 or Table S3.2) detected by the MoDiPath procedure to verify the effectiveness of our procedure.

Several of the automatically identified motifs listed in Table S4.1 or S4.2 (SLiMs shared by two or more than two species under study) are variations of the SKL<sub>S</sub> theme, where S represents a

**Table 1.** Number of motifs predicted in KEGG pathways.

Species	Total <sup>(a)</sup>			Significant SLiMs <sup>(b)</sup>			Novel SLiMs <sup>(c)</sup>		
	Total	MP	NMP	Total	MP	NMP	Tot	MP	NMP
<i>H.sapiens</i>	2097	836	1261	104	21	83	22	6	16
<i>M.musculus</i>	2094	882	1212	127	38	89	28	12	16
<i>R.norvegicus</i>	1863	809	1054	72	19	53	15	5	10
<i>D.melanogaster</i>	1391	632	759	35	5	30	4	0	4
<i>C.elegans</i>	1050	610	440	32	12	20	6	6	0
<i>E.coli</i>	933	733	200	11	10	1	2	1	1
<i>S.cerevisiae</i>	889	584	305	20	15	5	3	2	1

<sup>(a)</sup> Total number of motifs predicted by SIMFinder in KEGG pathways.

<sup>(b)</sup> number of significantly over-represented motifs in pathways with respect to the two reference datasets (hyper-geometric p-value < 3e-9, see Materials and Methods);

<sup>(c)</sup> number of significant motifs that are novel (hyper-geometric p-value < 3e-9, NormC < 0.7). MP: Metabolic pathways; NMP: Non-Metabolic Pathways.

doi:10.1371/journal.pone.0022270.t001

serine, K a lysine, L a leucine, and \$ indicates that true positive occurrences of the motif are found at the carboxy-terminal of proteins.

The SKL\$ motif significantly overlaps with the ELM TRG\_PTS1 motif (regular expression: ([SAPTC][KRHI][LMFI]\$]), which is annotated as a C-terminal signal interacting with the Pex5p protein to target proteins into the peroxisomal matrix, and is identical to a MnM motif annotated as Pex5 binding and associated to trafficking to Peroxisomes. Furthermore, Gould et al [35] identified the motif as a peroxisomal targeting signal in four unrelated peroxisomal proteins and both Miura et al [36] and Fujiki [37] found, more generally, that it functions as a topogenic signal in the translocation of proteins into peroxisomes. The signal needs to include the whole tripeptide sequence with a free alpha-COOH group at its carboxy terminus.

This motif is significantly over-represented in the Peroxisome KEGG pathway (KEGG ID: hsa04146) and specific (hyper-geometric p-value < 1.72e-11). Six proteins out of the sixty-nine belonging to this pathway share the motif. All of them are localized in the peroxisome, five of them participate to a fatty acid metabolic process and three of them have catalytic activity. Figure S1 shows the PROSITE [17] and Pfam [38] domain composition of these proteins together with the position of the SKL\$ motif in the sequence. Notably, the motif occurs in only 8 other sequences out of the 14,239 proteins of the non-redundant UniProt human dataset (filtered at the 40% sequence identity level). Of these, four are membrane or secreted proteins and therefore are likely to be false positives. The remaining four proteins are a peroxisomal acyl-coenzyme A oxidase 3 (UniProt O15254-1), a Lon protease homolog (Q86WA8), a peroxisomal leader peptide-processing protease (Q2T9J0), and a zinc-binding alcohol dehydrogenase domain-containing protein (Q8N4Q0). O15254-1 is a different isoform of O15254-2, a human protein, reported to belong to the hsa04146 KEGG pathway, that does not contain the motif and differs from O15254-1 for the lack of the last 75 C-term amino acids; it is not clear why O15254-2 was chosen for inclusion in the KEGG hsa04146 pathway; we argue that O15254-1 should be added to the KEGG hsa04146 pathway and the assignment of O15254-2 reassessed. Q86WA8 is annotated in UniProt for having the SKL\$ targeting motif and its cellular compartment is known to be the Peroxisome, but is not associated with any KEGG pathway. Q2T9J0 and Q8N4Q0 are peroxisomal proteins but they are neither annotated for having the motif nor associated with any KEGG pathway. We propose that Q2T9J0 and Q8N4Q0 use the SKL\$ motif as targeting signal to the peroxisome and suggest

that their inclusion, and that of Q86WA8, in the KEGG peroxisome pathway should be considered.

Another interesting motif that we automatically detected is WS.WS (Trp-Ser-any-Trp-Ser), which is specific for the Hematopoietic cell lineage pathway (KEGG ID: hsa04640) (hyper-geometric p-value < 3.10e-11). The motif was found in the analysis of both the 40% and 25% non-redundant sequence datasets and is present in 9 proteins out of the 79 belonging to the pathway, whereas it occurs in only 59 other sequences of the 40% non-redundant UniProt human dataset. Figure S2 shows the PROSITE [17] and Pfam [38] domain composition of the nine KEGG proteins together with the position of the WS.WS motif in the sequence: the motif is found at the C-terminal of the PROSITE FN3 domain in six cases and outside of the domain in three cases. This suggests that, at least in some of these proteins, the occurrence of the motif is not due to evolutionary conservation but rather to functional constraints. The WS.WS motif appears to be necessary for the binding activity of the erythropoietin receptor (EpoR), a member of the cytokine and growth factor receptor family. These proteins share conserved features in their extracellular and cytoplasmic domains presumably necessary for proper folding and thereby efficient intracellular transport and cell-surface receptor binding. Yoshimura et al [39] demonstrated that mutations in the motif of EpoR abolish processing, ligand binding, and activation of the receptor, while Schimmenti et al [40] showed that WS.WS is necessary for EpoR binding to Epo. For two (UniProt: P15509 and Q99062) out of the nine proteins hosting the motif, the crystal structure has been determined (PDB:3CXE [41] and 2D9Q [42]). In both cases, the motif instance is nicely found in an exposed loop of the protein structure (Figure 2).

The proteins belonging to the hematopoietic cell lineage pathway (KEGG ID: hsa04640) and sharing the motif all take part in two other pathways: Cytokine-cytokine receptor interaction (KEGG: hsa04060) and Jak-STAT signaling pathway (KEGG: hsa04630).

Our analysis also revealed that, out of the 59 other sequences of the non-redundant UniProt human dataset having the motif, 32 are likely to be false positives. The eighteen remaining proteins, that we estimated to be false negatives, have a similar molecular function (receptor activity) and a similar subcellular localization (membrane or secreted) of the true positives. Moreover, 16 of them are annotated in UniProt as having the functional motif, 13 are involved in both hsa04060 and hsa04630 KEGG pathways, one (Q14627) belongs to hsa04630, three (O75462, Q8IU8, Q8NI17)

**Table 2.** Number of KEGG pathways (total and with motifs).

Species	KEGG pathways <sup>(a)</sup>			Pathways with SLIMs <sup>(b)</sup>			Pathways with novel SLIMs <sup>(c)</sup>		
	Total	MP	NMP	Total	MP	NMP	Total	MP	NMP
<i>H.sapiens</i>	201	87	114	42	13	29	19	5	14
<i>M.musculus</i>	198	87	111	50	17	33	18	7	11
<i>R.norvegicus</i>	197	84	113	38	13	25	14	5	9
<i>D.melanogaster</i>	118	84	34	9	4	5	3	0	3
<i>Celegans</i>	117	82	35	15	9	6	4	4	0
<i>E.coli</i>	105	90	15	8	7	1	2	1	1
<i>S.cerevisiae</i>	92	70	22	11	9	2	2	1	1

<sup>(a)</sup>: Total number of KEGG pathways in each of the seven organisms under study.

<sup>(b)</sup>: Number of KEGG pathways for which at least one significant motif was found (hyper-geometric p-value < 3e-9, see Materials and Methods).

<sup>(c)</sup>: Number of KEGG pathways for which at least one statistically significant novel motif was found (i.e. a motif with no similarity to any known motif) (hyper-geometric p-value < 3e-9, NormC < 0.7). MP: Metabolic pathways; NMP: Non-Metabolic Pathways. doi:10.1371/journal.pone.0022270.t002

are included in KEGG but without pathway annotation, one (P40189) is not present in KEGG.

From the examples reported above and others reported in Tables S3.1 and S3.2, it is apparent that our automatic analysis can effectively discover biologically significant motifs and therefore that some of the novel ones (Tables S2.1 and S2.2), i.e. motifs not annotated in any other resource, might be interesting and worth investigating.

No matter how stringent are the statistical parameters used to identify significant hits, assessing the biological value of a short motif can only be achieved via experimental validation or by a carefully reviewing of the literature.

As an example of the usefulness of inspecting our proposed novel motifs and of the procedure that one can follow to gain confidence in the results, we illustrate here the case of the [FL]L.C.Y..A motif. This is conserved both in human (hsa04666) and mouse (mmu04666) Fc gamma R-mediated phagocytosis KEGG pathways. In the following we discuss the analysis of the human proteins sharing the motif, but the results are the same for the mouse proteins (data not shown).

The motif is present in 5/63 human proteins belonging to hsa04666: P42338, Q9Y217, Q13393, Q92608, O14939. P42338 is the catalytic subunit beta isoform of the phosphatidylinositol-4,5-bisphosphate 3-kinase, which phosphorylates several phosphoinositides [phosphatidylinositol (PtdIns), phosphatidylinositol 4-phosphate (PtdIns4P), phosphatidylinositol 4,5-bisphosphate (PtdIns(4,5)P2)] with a preference for PtdIns(4,5)P2. Phosphoinositides represent a small fraction of cellular phospholipids and are very important regulatory molecules utilized both as cellular membrane structural lipids and as precursors of multiple signalling molecules. Q9Y217 is a 1-phosphatidylinositol-3-phosphate 5 kinase. Q13393 and O14939 are phospholipases, which UniProt reports to be stimulated by PtdIns(4,5)P2 and PtdIns(3,4,5)P3 and by PtdIns(4,5)P2, respectively. Q92608 is a Dedicator of cytokinesis protein 2 (DOCK2). Interestingly, Nishikimi and colleagues [43] found that DOCK2 rapidly translocates to the plasma membrane in a PtdIns(3,4,5)-P3 dependent manner. In summary, all these proteins are involved in the interaction with phosphoinositides. By searching the motif in the whole set of human UniProt sequence, we found 9 additional occurrences in 9 different proteins. Three of them are isoforms of Q13393 and two are isoforms of O14939. Of the remaining four, one (O00329) is a PtdIns(4,5)P2 3-kinase catalytic subunit delta isoform, which is reported to be involved in the PtdIns phosphate biosynthesis, and

one (Q8TDW7) is the Protocadherin FAT-3. The molecular function of FAT-3 is not well known, however some authors [44,45] reported that the fat-3 gene acts in the same genetic pathway as synaptojanin, the main substrate of which in the brain is PtdIns(4,5)P2 and suggest that FAT-3 functions in the endocytic part of the synaptic vesicles recycling process. More specifically, Marza et al [45] found that the levels of PtdIns(4,5)P2 at release sites are increased in *Caenorhabditis elegans* fat-3 mutants lacking long-chain polyunsaturated fatty acids (LC-PUFAs), which would suggest that fat-3 influences the levels of PtdIns(4,5)P2 at release sites. For the remaining two proteins (O75976 and Q8NEZ3) we did not find any clue to deduce potential interactions with phosphoinositides and we cannot exclude that they are false positives. We also analysed the 58/63 hsa04666 proteins that do not have the [FL]L.C.Y..A motif. In this case, we automatically selected proteins that have at least one keyword related to phosphoinositides (e.g. PtdIns) in their UniProt annotation: we found ten of such proteins and inspected their sequences. In six of them, we found motifs that are similar, although not identical, to [FL]L.C.Y..A. For example, the P48736 sequence contains the subsequence FVYSCAGYCVVA which could be described by the [FL][LY]C.Y..A regular expression, a less specific version of the original expression. In the four remaining sequences, we did not find sub-sequences sufficiently similar to the identified motif.

In conclusion, our analysis suggests that the [FL]L.C.Y..A motif (and perhaps other related ones) is involved or participates in the recognition of phosphoinositides.

#### The MoDiPath Database and the Web Interface

The whole set of motifs identified by our procedure in the seven analysed organisms is stored in a MySQL database and made available to the scientific community through a Web Interface (<http://www.biocomputing.it/modipath>). Data are available for motifs identified in both the 40% and 25% datasets. The Web Interface has two main sections: "Search", for searching the MoDiPath database, and "Scan", for either searching motif matches in a protein sequence submitted by the user or for scanning the database with a user-defined regular expression. The MoDiPath database can be searched by KEGG pathway identifier, protein identifier (either UniProt or KEGG) and/or organism. The search by KEGG ID returns a table reporting the motif(s) associated with the input pathway. For each motif, the output provides the motif regular expression, indicates if the



**Table 3.** Number of motif representatives predicted in KEGG pathways.

Species	Total <sup>(a)</sup>			Significant SLiMs <sup>(b)</sup>			Novel SLiMs <sup>(c)</sup>		
	Total	MP	NMP	Total	MP	NMP	Tot	MP	NMP
<i>H.sapiens</i>	813	329	484	64	18	46	21	6	15
<i>M.musculus</i>	803	384	419	58	20	38	22	10	12
<i>R.norvegicus</i>	727	322	405	55	16	39	15	5	10
<i>D.melanogaster</i>	616	378	238	14	5	9	4	0	4
<i>Celegans</i>	513	307	206	20	11	9	5	5	0
<i>E.coli</i>	465	378	87	7	6	1	2	1	1
<i>S.cerevisiae</i>	502	336	166	16	13	3	2	1	1

<sup>(a)</sup>: Total number of motif representatives predicted by SLiMfinder in KEGG pathways;

<sup>(b)</sup>: number of significantly over-represented motif representatives in pathways with respect to the two reference datasets (hyper-geometric p-value < 3e-9, see Materials and Methods);

<sup>(c)</sup>: number of significant motif representatives that are novel (hyper-geometric p-value < 3e-9, NormC < 0.7). MP: Metabolic pathways; NMP: Non-Metabolic Pathways. doi:10.1371/journal.pone.0022270.t003

regular expression overlaps with at least one motif in another database (ELM, MnM, etc), reports the hyper-geometric p-value of the motif with respect to the SwissProt dataset (see Materials and Methods) and the fraction of proteins belonging to the pathway that contain the motif.

The system also provides further information on a specific motif, including

- the SlimFinder motif statistics;
- the sequence alignment and the list of proteins that both belong to the pathway AND contain the motif;
- the motif cross-reference to other databases of motifs;
- the list of GO terms shared by the proteins matching the motif;
- PROSITE [17] and Pfam [38] domains shared by the protein sequences matching the motif;
- the exact sequence of the motif;
- the starting and ending position of the match in the protein sequence;
- the evolutionary conservation score;
- the PDB ID (if available).
- access to the STRING database [46] that provides an interaction map specific for the proteins of the pathway sharing the motif.

Figure 3 shows a screenshot with the information provided by MoDiPath for the WS.WS motif, which is specific for the



**Figure 2.** The crystal structure of the human granulocyte colony-stimulating factor (G-CSF) receptor. The structure of the G-CSF receptor (PDB:2D9Q) [42] is reported in orange. Residues corresponding to the WS.WS motif (residues 295–299) are shown in blue. doi:10.1371/journal.pone.0022270.g002

hsa04640 KEGG pathway. For each protein sharing the motif, a page containing functional and structural details is provided. In particular, if the protein is of known structure, the position of the matching sub-sequence is displayed in the context of its three-dimensional structure.

If the initial search is performed with a protein ID, the list of pathways including the query protein and, for each pathway, the list of motifs matching the protein, if any, can be retrieved.

A search by organism returns the list of KEGG pathways for which at least one statistically significant motif has been found in the query organism. Each pathway is linked to the complete list of its motifs.

Finally, for each motif it is possible to download, explore and edit the whole pathway map corresponding to a selected motif using KGML-ED [47], a Web Java start program downloadable through the MoDiPath Web Interface. In each pathway map, proteins containing the motif are conveniently highlighted.

The implementation of the complete system can also be downloaded and installed locally to analyse other organisms of interest or to use definition of pathways provided by other resources such as PANTHER [48], REACTOME [49], or EcoCyc and MetaCyc databases [50].

## Discussion

The discovery of linear motifs is a difficult task that usually requires the identification of a set of non-homologous proteins sharing a common functional feature (e.g., an interaction partner or a cellular compartment). Many algorithms for motif discovery are nowadays available and appropriate statistics have been developed for estimating the effectiveness of a motif for function prediction. However, several challenging aspects still remain, for example one needs to identify appropriate sets of non-homologous proteins sharing a functional feature and associate the appropriate biological function to newly discovered motifs. The two issues are of course strictly related: for example, if one were able to identify a set of proteins that are targeted to the same cellular compartment, a motif significantly over represented in their sequences would be likely to be a targeting signal to that compartment.

This is the idea that inspired several works in the field, such the one of Neduva et al, aimed at discovering motifs that mediate protein-protein interaction networks [51].

Restricting the analysis to non-homologous proteins is relevant to avoid detecting general sequence homology features instead of

### Pathway ID: [hsa04640](#)

### Pathway Name: Hematopoietic cell lineage

Motif regular expression	Match in DB(s)	Hyper.G	Fraction
WS.WS	±	3.10e-11	9/79

Sig	SeqNum(s)	UPNum	AANum	MotNum	Rank	IC	Occ	Support	UP	ExpUP	Prob	Cloud	CloudSeq	CloudUP
6.94e-05	79	54	39316	69	9	4.0	9.0	9	7	0.1	1.61e-11	1	18	14

(9) WS.WS

Kegg ID	Overlapping Domain ID	Sequence	Start-End	Conservation score	UniProt Ac	UniProt ID	PDB ID	Show details
<a href="#">hsa:3581</a>	<a href="#">ps:HEMATOPO_REC_S_F1</a>	WSEWS	245-249	1	<a href="#">Q01113</a>	<a href="#">IL9R_HUMAN</a>	NONE	Show
<a href="#">hsa:1438</a>		WSSWS	306-310	0.48	<a href="#">P15509</a>	<a href="#">CSP2R_HUMAN</a>	<a href="#">3CXE</a>	Show

Term ID	Term name	Fraction
<b>Cellular Component</b>		
GO:0016020	membrane	9/9
GO:0016021	integral to membrane	9/9
GO:0005576	extracellular region	7/9
GO:0005886	plasma membrane	7/9
GO:0005887	integral to plasma membrane	6/9
<b>Biological Process</b>		
GO:0007165	signal transduction	6/9

Database	Motif ID
ELM	MOD_CK1
ELM	MOD_CMANS
MnM	PPSSTX0000
MnM	PPSSTX0000
MnM	PBMAP20000
MnM	PRMETM0000
MnM	PRMMAN0000
MnM	ELMERR0000
phosphomotif	KIN_ST030

Kegg ID	Domain ID	hsa:3568	hsa:3581	hsa:3590	hsa:1438	hsa:3568
<b>Pfam Domains</b>						
BAH						
DUF2417		●				
DUF3184					●	
EpoR_Lig-bind		●		●		
I-set				●		

**Figure 3. The information provided by MoDiPath for the hsa04640 KEGG pathway.** (a) First column: the SLiM regular expression; Second column: a '+' is reported if the motif overlaps to a similar motif in other databases (the list of which is shown by moving the mouse over the '+'); Third column: the hyper-geometric p-value of the number of motif hits in the hsa04640 pathway compared to the number of motif hits in the SwissProt database; Fourth column: The fraction of proteins in the hsa04640 pathway that contain the WS.WS motif (b) Multiple sequence alignment of the hsa04640 pathway proteins containing the WS.WS motif. (c) Information about each of the hsa04640 proteins containing the WS.WS motif. Clicking on the 'Show' button provides more detailed information, including the protein structure visualization with the motif hit(s) highlighted. (d) List of motif overlap(s) to similar motifs in other databases; the last column reports the CompaMotif [32] similarity score (NormIC). (e) GO terms shared by the hsa04640 pathway proteins that have the motif; the last column reports the fraction of the proteins hosting the motif that share a GO term. doi:10.1371/journal.pone.0022270.g003

genuine functional motifs. Even though functional motifs can also be found in evolutionary related proteins, the most interesting ones are represented by cases of convergent evolution. However, the latter are rare and difficult to discover, especially at the level of the protein sequence. One possible approach to identify motifs arising independently during evolution, consists, on one hand, in using non-homologous sequences and, on the other, in filtering out motifs occurring in similar (e.g. Pfam) domains. The MoDiPath database only collects motifs identified in non-redundant sets of proteins and annotates motif matching proteins for the presence of Pfam and PROSITE domains. This facility does not ensure that every discovered motif will be a case of convergent evolution, but can help users identify those that are likely to be relicts of common descent with no specific functional properties.

Here, we focused on functional features typical of metabolic and signaling pathways. Pathway functional features can be of different types: they could be related to the interaction with the same metabolite or its derivatives, or pertain to specific cellular compartments, or arise, for example, from the interaction with recurring signaling modular domains (SH2, SH3, WW, PDZ, etc).

We used this strategy to explore all proteins of seven organisms assigned to KEGG pathways and identified a number of potentially biologically significant motifs that represent a valid starting point for further computational and experimental functional investigation.

The methodology is reliable, as demonstrated by the fact that we can automatically re-discover known motifs, for example the targeting peroxisome signal SLK\$ or the WS.WS motif necessary for processing, ligand binding and activation of receptors specific for the hematopoietic cell lineage pathway but also taking part in two related pathways: the cytokin-cytokine receptor interaction pathway and the Jak-STAT signaling pathway.

The procedure is also effective in detecting novel motifs. As an example we described here the analysis of one of them ([FL]L.C..Y..A) for which no functional annotation is available, and found that it is likely to be involved in the recognition of phosphoinositides.

We hope that MoDiPath, its associated database as well as the list of motifs that we provide here will contribute to speed up the discovery of novel motifs and will constitute a useful resource for the life scientists.

## Materials and Methods

### Motif discovery procedure

We used the KEGG (Kyoto Encyclopedia of Genes and Genomes) Pathway database as the source of pathway information. In this resource, proteins from 1173 different species (release of March 2010) [27] are clustered in pathways. Each pathway represents functional aspects of a biological system, and involves a specific protein list, graphically represented as a network of connected proteins. The number of pathways depends on the species (Table 2).

Pre-computed data presently associated with MoDiPath are available for seven species (*H.sapiens*, *R.norvegicus*, *M.musculus*, *D.melanogaster*, *C.elegans*, *S.cerevisiae*, *E.coli*).

For each KEGG pathway, we collected all protein sequences and, in order to only retain unrelated proteins, used CD-HIT (last release 4.0 beta) [30] to remove redundancy at the 40% as well as at the 25% identity level. Each pathway protein list was analysed by SlimFinder, one of the best performing tools for linear motif detection [25]. In SlimFinder the term SLiM is used to mean short (generally less than 10 residues), linear (i.e. made up of adjacent residues in the primary sequence) true functional motif. SLiMs, which are encoded by regular expressions, are composed by defined amino acid positions often separated by wildcards (which represent positions that can be occupied by any amino acid). Defined position can be fixed (only one amino acid type is permitted) or degenerate (more than one amino acid type is permitted). The number of defined positions and of wildcards can be either fixed or variable.

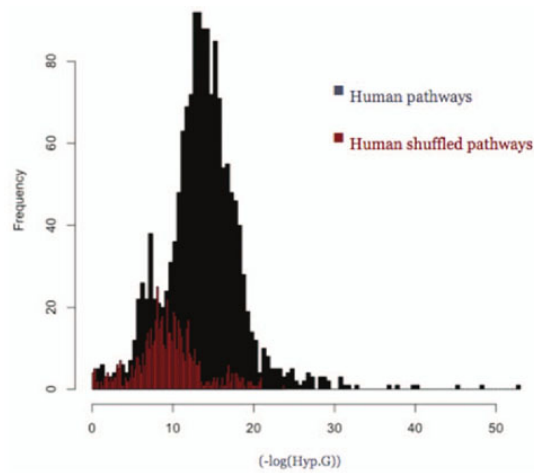
SlimFinder is a software package that implements two different algorithms, SlimBuild and SlimChance, and offers a number of input masking options, which can be used to restrict the analysis to specific parts of the proteins, such as disordered or low complexity regions. SlimBuild builds motifs by first combining pairs of residues into longer patterns and subsequently incorporating amino acid degeneracy and/or variable length wildcards, until the SLiM matches the desired number of unrelated sequences. SlimChance deals with the probability that a motif occurs in a sequence dataset by chance and determines a score indicating how unlikely a given motif is compared to other motifs in a dataset.

The input of SlimFinder is a user-defined set of sequences, plus a number of options such as the BLAST e-value threshold to be used to identify which input proteins are related to which other input proteins, the minimum number of unrelated proteins that should contain the motifs, the maximum number of defined positions in a motif, the maximum number of wildcard positions, disorder masking, etc.

SlimFinder was run locally with default parameters except for the disorder masking option, which was deactivated. We retained the subset of top significant motifs with a very high probability of significance (SlimChance probcut = 0.99).

The statistical assessment of a motif specificity for a given pathway was obtained by comparing the number of the motif occurrences in the proteins belonging to the pathway with the number of occurrences in two reference datasets: 1) all UniProt proteins (from the same organism) and 2) all the proteins included in KEGG. Since proteins belonging to a KEGG pathway are contained in both reference datasets, the hyper-geometric p-value was used to assess the motif specificity, i.e. to assess whether it is observed more frequently in the KEGG pathway than expected by chance given its frequency in each of the two reference datasets.

In order to choose an un-biased hyper-geometric p-value threshold, we had to take into account the KEGG pathway peculiar composition, which is clearly not random. To this aim, we



**Figure 4. Motif occurrence Hyper-geometric distribution.** Hyper-geometric  $p$ -value distribution for the number of motif occurrences in true (black) and reshuffled (red) KEGG pathways with respect to the number of motif occurrences in the UniProt dataset for *H.sapiens*. The  $p$ -value =  $3e-9$  approximately corresponds to a false discovery rate of 10%. doi:10.1371/journal.pone.0022270.g004

built random pathways by reshuffling the proteins of each pathway with proteins belonging to other pathways, leaving the number of proteins per pathway unmodified. Next, we plotted the hyper-geometric  $p$ -value distribution of motif occurrences in the random datasets with respect to their occurrences in the UniProt dataset and compared it to the corresponding distribution for the true datasets (Figure 4). We estimated that the hyper-geometric  $p$ -value that better discriminates between true and false positives (random) is  $3e-9$ , which corresponds to a false discovery rate (FDR) lower than 10%. The procedure was repeated ten times for *H.sapiens* producing essentially the same result. The result was the same when all human proteins of SwissProt were used for reshuffling (data not shown).

#### Motif-motif comparison

The CompariMotif software [32] was used to compare predicted motifs to similar motifs annotated in other databases (ELM [18], MnM [19], PhosphoMotif Finder [33]), a set of SLiMs extracted from the literature, and predicted SLiMs from Neduva & Russell [1]. The software takes as input two lists of motifs and returns a set of motif pairs associated with a similarity score (Normal IC), which ranges between 0.0 (weak similarity) and 1.0 (strong similarity).

CompariMotif uses a sliding window to compare every possible alignment between two motifs (represented as regular expressions). Two aligned positions are considered a mismatch if they have no amino acid in common amino (in which case the motif pair is rejected). Each compared position is scored according to its information content:  $IC_i = -\log_2(f_{i_a})$ , where  $IC_i$  is the information content for position  $i$ ,  $f_{i_a}$  is the summed frequency for the amino acids at position  $i$ , and  $N = 20$ .  $IC_i$  is a modification of the

Shannon's Information Content algorithm [52] where wildcards have score 0, fixed positions have score 1, and ambiguous positions have scores between 0 and 1. The  $IC_m$  of a match is the sum of the component  $IC_i$  values. A sliding window will produce several matches and the best match is taken as the one with the best overall  $IC_m$ . In order to make the score independent from the length and degeneracy of the matching motifs, a final normalized IC (Norm IC) score is calculated by dividing the  $IC_m$  by the lower IC value for the two motifs. Pairs of motifs with Norm IC = 0 are clearly dissimilar and pairs of motifs with Norm IC = 1 are highly similar, however, a cut-off must be set for pairs of motifs with intermediate Norm IC values in order to discriminate between true and false matches. The choice of such cut-off is arbitrary and depends on the empirical observation of compared motifs (RJ Edwards, personal communication).

Based on the analysis of Norm IC scores for MoDiPath pairs of compared motifs, we considered two SLiMs to be similar if their Normal IC  $> 0.7$ .

#### Supporting Information

**Supporting Information S1 Supplementary motif and pathway statistics.** The file contains the same data of Table 1, 2, and 3 (main text) calculated for the 25% non-redundant sequence dataset. Moreover, it reports statistics on motifs occurring in disordered and loop regions. It is organized in three sections as follows: 1) Motif and pathway statistics calculated for the 25% non-redundant sequence dataset. 2) Statistics of motifs occurring in disordered regions of proteins (calculated for both the 40% and 25% datasets). 3) Statistics on motifs occurring in loop regions (calculated for both the 40% and 25% datasets). A motif is



assigned to a loop (disordered) region if at least 50% of the residues belonging to the motif true positive matches are in loop (disordered) regions, respectively.  
(DOC)

**Table S1 Total number of proteins belonging to the pathways under study and number of motifs per pathway.** Table S1.1: Data obtained from the analysis of the 25% non-redundant sequence dataset. Table S1.2: Data obtained from the analysis of the 40% non-redundant sequence dataset.  
(XLS)

**Table S2 List of novel motifs.** Table S2.1: List of novel motifs obtained by restricting the analysis to the 25% non-redundant sequence dataset. Table S2.2 – List of novel motifs obtained by restricting the analysis to the 40% non-redundant sequence dataset.  
(XLS)

**Table S3 List of known motifs.** Table S3.1: List of known motifs obtained by restricting the analysis to the 25% non-redundant sequence dataset. Table S3.2: List of known motifs obtained by restricting the analysis to the 40% non-redundant sequence dataset.  
(XLS)

**Table S4 List of motifs shared by two or more of the species under study.** Table S4.1: List of motifs shared by two or more of the species under study in the 25% sequence non-redundant dataset. Table S4.2: List of motifs shared by two or more of the species under study in the 40% non-redundant sequence dataset.  
(XLS)

## References

- Nedeva V, Russell RB (2005) Linear motifs: evolutionary interaction switches. *FEBS Lett* 579: 3342–3345.
- Geol A, Chaitry-aramonri A, Santonico E, Sacco R, Castagnoli L, et al. (2007) DOMINO: a database of domain-peptide interactions. *Nucleic Acids Res* 35: D557–560.
- Nedeva V, Russell RB (2006) Peptides mediating interaction networks: new leads at last. *Curr Opin Biotechnol* 17: 465–471.
- Russo T, Faraonio R, Minopoli G, De Candia P, De Renzi S, et al. (1998) Fe65 and the protein network centered around the cytosolic domain of the Alzheimer's beta-amyloid precursor protein. *FEBS Lett* 434: 1–7.
- Pasani LA, Bedford MT, Faber PW, McGinnis KM, Sharp AH, et al. (2000) Huntington's WW domain partners in Huntington's disease post-mortem brain fulfill genetic criteria for direct involvement in Huntington's disease pathogenesis. *Hum Mol Genet* 9: 2175–2182.
- Huang X, Poy F, Zhang R, Joachimiak A, Sudd M, et al. (2000) Structure of a WW domain containing fragment of dystrophin in complex with beta-dystroglycan. *Nat Struct Biol* 7: 634–638.
- Marti M, Good RT, Rug M, Kneuper E, Cowman AF (2004) Targeting malaria virulence and remodeling proteins to the host erythrocyte. *Science* 306: 1930–1933.
- Hiller NL, Bhattacharjee S, van Ooij C, Lidios K, Harrison T, et al. (2004) A host-targeting signal in virulence proteins reveals a secretome in malarial infection. *Science* 306: 1934–1937.
- Furmanek A, Hofsteenge J (2000) Protein C-mannosylation: facts and questions. *Acta Biochim Pol* 47: 781–789.
- Tong AH, Drees B, Nardelli G, Bader GD, Brannetti B, et al. (2002) A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science* 295: 321–324.
- Cesareni G, Panni S, Nardelli G, Castagnoli L (2002) Can we infer peptide recognition specificity mediated by SH3 domains? *FEBS Lett* 513: 38–44.
- Munro S, Pelham HR (1987) A C-terminal signal prevents secretion of luminal ER proteins. *Cell* 48: 899–907.
- Hu H, Columbus J, Zhang Y, Wu D, Lian L, et al. (2004) A map of WW domain family interactions. *Proteomics* 4: 643–655.
- Miller ML, Jensen IJ, Diella F, Jørgensen C, Timi M, et al. (2008) Linear motif atlas for phosphorylation-dependent signaling. *Sci Signal* 1: ra2.
- Fuxreiter M, Tompa P, Simon I (2007) Local structural disorder imparts plasticity on linear motifs. *Bioinformatics* 23: 950–956.
- Diella F, Haslam N, Chica C, Budd A, Michael S, et al. (2008) Understanding eukaryotic linear motifs and their role in cell signaling and regulation. *Front Biosci* 13: 6580–6603.
- Hulo N, Bairoch A, Bullard V, Cerutti L, De Castro E, et al. (2006) The PROSITE database. *Nucleic Acids Res* 34: D227–230.
- Gould CM, Diella F, Via A, Puntervoll P, Gemund C, et al. ELM: the status of the 2010 eukaryotic linear motif resource. *Nucleic Acids Res* 38: D167–180.
- Balla S, Thapar V, Verma S, Luong T, Faghri T, et al. (2006) Minimoto Miner: a tool for investigating protein function. *Nat Methods* 3: 175–177.
- Bailey TL, Boden M, Buske FA, Friti M, Grant CE, et al. (2009) MEME-SUITE: tools for motif discovery and searching. *Nucleic Acids Res* 37: W202–208.
- Nedeva V, Russell RB (2006) DILMOT: discovery of linear motifs in proteins. *Nucleic Acids Res* 34: W350–353.
- Rigoutsos I, Floratos A (1998) Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm. *Bioinformatics* 14: 55–67.
- Marcattili P, Bassotti G, Tramontano A (2008) The MoVIN server for the analysis of protein interaction networks. *BMC Bioinformatics* 9 Suppl 2: S11.
- Davey NE, Shields DC, Edwards RJ (2006) SLMDic: short, linear motif discovery, correcting for common evolutionary descent. *Nucleic Acids Res* 34: 3546–3554.
- Davey NE, Haslam NJ, Shields DC, Edwards RJ (2006) SLMFinder: a web server to find novel, significantly over-represented, short protein motifs. *Nucleic Acids Res* 34 Suppl: W534–539.
- Tu BP, Kudlicki A, Rowicka M, McKnight SL (2005) Logic of the yeast metabolic cycle: temporal compartmentalization of cellular processes. *Science* 310: 1152–1158.
- Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 28: 27–30.
- Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, et al. (1997) CATH – a hierarchical classification of protein domain structures. *Structure* 5: 1093–1108.
- Chothia C, Lesk AM (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J* 5: 823–826.
- Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22: 1658–1659.
- Consortium U (2010) The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res* 38: D142–148.
- Edwards RJ, Davey NE, Shields DC (2008) CompariMotif: quick and easy comparisons of sequence motifs. *Bioinformatics* 24: 1307–1309.
- Amamchy R, Perisawany B, Mathivanan S, Reddy R, Tattikonda SG, et al. (2007) A curated compendium of phosphorylation motifs. *Nat Biotechnol* 25: 285–286.
- Chica C, Labarga A, Gould CM, Lopez R, Gibson TJ (2008) A tree-based conservation scoring method for short linear motifs in multiple alignments of protein sequences. *BMC Bioinformatics* 9: 229.
- Gould SJ, Keller GA, Subramani S (1988) Identification of peroxisomal targeting signals located at the carboxy terminus of four peroxisomal proteins. *J Cell Biol* 107: 897–905.

36. Miura S, Kawaya-Arai I, Mori H, Miyazawa S, Osumi T, et al. (1992) Carboxyl-terminal consensus Ser-Lys-Leu-related tripeptide of peroxisomal proteins functions in vitro as a minimal peroxisome-targeting signal. *J Biol Chem* 267: 14405-14411.
37. Fujiki Y (1992) [Biogenesis of peroxisome-targeting signal and peroxisome assembly factor]. *No To Hattatsu* 24: 181-183.
38. Finn RD, Mistry J, Tate J, Coghill P, Heeger A, et al. (2010) The Pfam protein families database. *Nucleic Acids Res* 38: D211-222.
39. Yoshimura A, Zimmers T, Neumann D, Longmore G, Yoshimura Y, et al. (1992) Mutations in the Trp-Ser-X-Trp-Ser motif of the erythropoietin receptor abolish processing, ligand binding, and activation of the receptor. *J Biol Chem* 267: 11619-11625.
40. Schimmenti LA, Blecher G, Harris KW, Winkelmann JC (1995) Localization of an essential ligand binding determinant of the human erythropoietin receptor to a domain N-terminal to the WSKWS motif: implications for soluble receptor function. *Exp Hematol* 23: 1341-1346.
41. Hansen G, Hercus TR, McClure BJ, Sosanski FC, Dostore M, et al. (2006) The structure of the GM-CSF receptor complex reveals a distinct mode of cytokine receptor activation. *Cell* 134: 496-507.
42. Tamada T, Honjo E, Maeda Y, Okamoto T, Ishibashi M, et al. (2006) Homodimeric cross-over structure of the human granulocyte colony-stimulating factor (G-CSF) receptor signaling complex. *Proc Natl Acad Sci U S A* 103: 3135-3140.
43. Nishikimi A, Fukushima H, Su W, Hongu T, Takasuga S, et al. (2009) Sequential regulation of DOCK2 dynamics by two phospholipids during neutrophil chemotaxis. *Science* 324: 384-387.
44. Lesa GM, Palfreyman M, Hall DH, Clandinin MT, Rudolph C, et al. (2003) Long chain polyunsaturated fatty acids are required for efficient neurotransmission in *C. elegans*. *J Cell Sci* 116: 4965-4975.
45. Marza E, Long T, Sairadi A, Sumakovic M, Eimer S, et al. (2008) Polyunsaturated fatty acids influence synaptotagmin localization to regulate synaptic vesicle recycling. *Mol Biol Cell* 19: 833-842.
46. Jensen IJ, Kuhn M, Stark M, Chaffron S, Creevey C, et al. (2009) STRING 8—a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res* 37: D412-416.
47. Klukas C, Schreiber F (2007) Dynamic exploration and editing of KEGG pathway diagrams. *Bioinformatics* 23: 344-350.
48. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, et al. (2003) PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res* 13: 2129-2141.
49. Matthews I, Gopinath G, Gillespie M, Caudy M, Croft D, et al. (2009) Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res* 37: D619-622.
50. Karp PD, Riley M, Saier M, Paulsen IT, Paley SM, et al. (2000) The EcoCyc and MetaCyc databases. *Nucleic Acids Res* 28: 56-59.
51. Nedra V, Linding R, Su-Angrand I, Stark A, de Masci F, et al. (2005) Systematic discovery of new recognition peptides mediating protein interaction networks. *PLoS Biol* 3: e105.
52. Shannon CE (1997) The mathematical theory of communication. 1963. *MD Comput* 14: 306-317.

**Paper III: The interaction network of the 14-3-3 protein in the ancient protozoan parasite *Giardia duodenalis***

Journal of Proteome Research

**The interaction network of the 14-3-3 protein in the ancient protozoan parasite *Giardia duodenalis***

Journal:	<i>Journal of Proteome Research</i>
Manuscript ID:	pr-2011-00742g.R1
Manuscript Type:	Article
Date Submitted by the Author:	22-Aug-2011
Complete List of Authors:	Lalle, Marco; Istituto Superiore di Sanità Camerini, Serena; Istituto Superiore di Sanità Cecchetti, Serena; Istituto Superiore di Sanità Sayadi, Ahmed; University of Rome "Sapienza", Department of Biochemical Sciences "A. Rossi-Fanelli", Crescenzi, Marco; Istituto Superiore di Sanità Pozio, Edoardo; Istituto Superiore di Sanità

SCHOLARONE™  
Manuscripts

ACS Paragon Plus Environment



1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

**The interaction network of the 14-3-3 protein in the ancient protozoan parasite *Giardia duodenalis***

**Marco Lalle<sup>1\*</sup>, Serena Camerini<sup>2</sup>, Serena Cecchetti<sup>2</sup>, Ahmed Sayadi<sup>3</sup>, Marco Crescenzi<sup>2</sup>, Edoardo Pozio<sup>1</sup>**

<sup>1</sup>Department of Infectious, Parasitic and Immunomediated Diseases, Istituto Superiore di Sanità, 00161

Rome, Italy; <sup>2</sup>Department of Cell Biology and Neurosciences, Istituto Superiore di Sanità, 00161 Rome, Italy;

<sup>3</sup>Department of Biochemical Sciences "A. Rossi-Fanelli", University of Rome "Sapienza", 00185 Rome, Italy.

\*Corresponding author: Marco Lalle PhD, Department of Infectious, Parasitic and Immunomediated

Diseases, Istituto Superiore di Sanità, Viale Regina Elena 299, 00161 Rome, Italy. Tel. +39 06 4990 2670, Fax

+039 06 4990 3561, E-mail: [marco.lalle@iss.it](mailto:marco.lalle@iss.it)

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60**Abstract:**

The 14-3-3s, are a family of eukaryotic phosphoserine/phosphothreonine binding proteins that play pivotal roles as regulator of multiple cellular processes. The flagellated protozoan parasite *Giardia duodenalis* (syn. *lamblia* or *intestinalis*), the causing agent of giardiasis, is a valuable simplified eukaryotic model and harbors a single 14-3-3 isoform (g14-3-3) directly involved in the parasite differentiation into cyst. To define the role of g14-3-3 we investigated the protein interactome. A transgenic *G. duodenalis* strain was engineered to express a FLAG-tagged g14-3-3 under its own promoter. Affinity chromatography coupled with tandem mass spectrometry analysis has been used to purify and identify FLAG-g14-3-3-associated proteins from trophozoites and encysting parasites. A total of 314 putative g14-3-3 interaction partners were identified, including proteins involved in DNA replication, energy metabolism, cytoskeleton organization, and protein trafficking. Some interactions were observed to occur uniquely in one stage, while others were shared. Furthermore, the interaction of g14-3-3 with the giardial homolog of the CDC7 protein kinase (gCDC7) was characterized, leading to the identification of a multiprotein complex containing g14-3-3, gCDC7 and a newly identified and highly divergent homolog of DBF4, the putative regulatory subunit of gCDC7. We discuss the relevance of g14-3-3 interactions in *G. duodenalis* biology.

**Keywords:** g14-3-3 protein, protein-protein interaction, *Giardia duodenalis*, encystation, gCDC7

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

## 1. Introduction

The fine tuning of phosphorylation/dephosphorylation status of proteins is widely used by eukaryotic cells to regulate multiple cellular processes. In this scenario, despite the fundamental activity of different protein kinases and phosphatases, a key role is played by the dimeric 14-3-3, a family of highly conserved proteins that bind to specific Ser/Thr phosphorylated sites on target proteins. One or two 14-3-3 isoforms are present in unicellular organisms whereas metazoans harbor several isoforms thus allowing the formation of homo- and heterodimers (e.g. seven 14-3-3-coding genes in *Homo sapiens* and fifteen in *Arabidopsis thaliana*).<sup>1, 2</sup> The general 14-3-3 structure is well conserved between the different isoforms. Each monomer consists of nine  $\alpha$ -helices, organized in a cup-like shape, with the dimerization domain located at the N-terminal portion.<sup>3</sup> The interaction of 14-3-3s with the target proteins is mediated by conserved residue located in an amphipathic groove of each monomer and requires specific binding motifs on the targets. Three general consensus sequences for 14-3-3 binding have been defined: the mode-1 motif R[S/Ar][+Ar]p(S/T)[L/E/A/M]P (where p(S/T) are phosphorylated serine or threonine residues, Ar indicates an aromatic residue and + a basic residue), the mode-2 motif RX[S/Ar][+]p(S/T)[L/E/A/M]P<sup>4,5</sup> and the mode-3 motif -pS/pT X1-2-COOH,<sup>6</sup> in which the pSer or pThr is the penultimate residue in the C-terminal tail of the target protein. Nevertheless, the interaction with 14-3-3s may occur with other phosphopeptide sequences and/or through certain non-phosphorylated sequences.<sup>2,7,8</sup> 14-3-3 dimers are highly rigid structures and binding can induce conformational changes in protein ligands affecting their functions by inter- and intracompartamental sequestration, activation/inactivation of enzymatic activity, and, in few cases, by promotion/inhibition of protein-protein interaction. Focused works and large scale proteomic studies, from yeast to humans, have lead to the identification of hundreds of intracellular target proteins including enzymes and structural components of metabolism, intra- and extracellular protein trafficking, cytoskeleton, DNA replication, transcription, translation, cell cycle regulation and signal transduction pathways.<sup>9,20</sup> Understanding the roles of 14-3-3/targets interaction is progressively contributing to elucidate fine regulatory mechanisms in a wide range of eukaryotic processes. *Giardia duodenalis* (syn. *lamblia* or *intestinalis*) is a flagellated protozoan that parasitizes the upper part of the small intestine of

1  
2  
3 mammals, including humans, thus causing giardiasis, the most common non-bacterial and non-viral  
4  
5 diarrheal disease, estimated to affect 280 million people each year.<sup>21</sup> *G. duodenalis* (here referred as  
6  
7 *Giardia*) is a valuable eukaryotic model due to its minimalistic genomic and cellular organization.<sup>22</sup> This  
8  
9 parasite has a simple two stages life cycle, totally reproducible in the laboratory, that consists of: i) the  
10  
11 binucleated trophozoite, that replicates and colonizes the host intestine and ii) the tetranucleated cyst, the  
12  
13 infective and resistant stage, able to survive in the external environment. The infection starts by ingesting  
14  
15 cysts that undergo excystation into trophozoites in the proximal small intestine. After exposure to biliary  
16  
17 fluid, in the jejunum, some of the trophozoites undergo a cell differentiation process termed encystation  
18  
19 and form cysts that are spread in the environment with the faeces. *Giardia* harbors a single 14-3-3 isoform  
20  
21 (g14-3-3) that binds to target proteins via the conserved residues located in the amphipathic groove.<sup>23</sup>  
22  
23 Intriguingly, g14-3-3 is constitutively phosphorylated on Thr214, whereas other 14-3-3s are phosphorylated  
24  
25 only in certain conditions<sup>2</sup> and, uniquely, polyglycylated on Glu246 in a stage-dependent manner.<sup>23,24</sup>  
26  
27 Mutations of g14-3-3 and/or alterations of the level of its post-translational modifications influence the  
28  
29 ability of the parasite to develop into cyst, potentially reducing the parasite spread in the environment.<sup>24,25</sup>  
30  
31 Due to the evident role of g14-3-3 in parasite developing processes, the identification of the 14-3-3  
32  
33 interacting partners will provide novel information in the biology of *Giardia*. To survey the scope of 14-3-3  
34  
35 interactors, we performed a MS-based proteomic analysis of *in vivo* affinity-purified g14-3-3 complexes  
36  
37 from *Giardia* trophozoites and from parasites undergoing encystation.  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

## 2. Materials and Methods

### 2.1 Parasite cultivation and transfection

Trophozoites of the *G. duodenalis* WB-C6 were axenically grown and induced to encystation as previously described.<sup>24</sup> Transgenic *Giardia* lines were generated by electroporation in the presence of 15 µg of plasmid DNA and selection in the presence of 100 µM puromycin (Invivogen, Toulouse, France).<sup>24</sup>

### 2.2 Nucleic acid isolation

Genomic DNA was isolated from 10<sup>9</sup> trophozoites of *Giardia* WB-C6 using the phenol/chloroform extraction method. Total RNA was extracted from 10<sup>7</sup> trophozoites, or encysting parasites, using the RNeasy mini kit (Qiagen, Hilden, Germany), following the manufacturer's instructions. Plasmid DNA was isolated from bacteria using the QIAprep kit (Qiagen).

### 2.3 Vectors construction

*Escherichia coli* JM109 competent cells were used for vector manipulation and propagation. A 100 bp sequence encompassing the putative *g14-3-3* gene promoter region was PCR amplified from the genomic DNA of *Giardia* WB-C6 using the designed primers 14promF (5'-AAGCTTCAACGTAGCTGTACAGTGTC-3', the HindIII site is underlined) and 14promR (5'-CCATGGAGGTTTTTATTTAAGCTGCTG-3', the NcoI site is underlined). PCR reaction was performed in a final volume of 50 µl using 50 µl of 10X buffer containing 20 mM MgCl<sub>2</sub> (Takara Holdings Inc.; Kyoto, Japan), 50 µM dNTPs (Takara) 20 pmols of each primer and 1.25 units of ExTaq (Takara). Reactions were performed on a T-Personal Thermocycler (Biometra Corporation, Göttingen, Germany). Amplification conditions were: one cycle at 95°C for 5 min; 30 cycles at 95°C for 30 sec, 55°C for 30 sec and 72°C for 30 sec; and one cycle at 72°C for 7 min. The PCR fragment was cloned at the 5' end of the FLAG-tagged *g14-3-3* coding sequence in the HindIII/NcoI-digested *pya*-FLAG14-3-3 vector<sup>24</sup> replacing the promoter of the glutamate dehydrogenase (*gdh*) gene. The obtained plasmid was designated as "*pya*-P14\_FLAG14-3-3". A short linker encoding the FLAG epitope and containing a BamHI and PspOMI site for cloning of PCR fragments in frame with the FLAG (5'-ACATGTTGGATTATAAGGATGATGATAAG**GGATCC**GGGCCCAAA**TGATCA**-3', the PciI site is underlined, the BamHI site is in bold, the PspOMI site is in italic, and the BclI site is in bold and

1  
2  
3 underlined) was cloned in the NcoI/BamHI-digested PtubApaH7-HApac vector,<sup>26</sup> instead of the VSPH7 gene  
4 and of 3xHA tags coding sequence, to obtain the pTUB-FLAGpac vector. The full length coding sequence of  
5 the giardial *cdc7* homolog gene (GL50803\_112076) has been PCR amplified from *G. duodenalis* WB-C6 clone  
6 genomic DNA using the primers CDC7forw1 (5'-GGATCCATGACCCGACGCCACCAGGCC-3', the BamHI site is  
7 underlined) and CDC7rev (5'-GGGCCCTCGAGTTAGAAGTATATCTCGGCAC-3', the PspOMPI site is  
8 underlined). A central portion of the gene (CDC7-II, 1398bp), encompassing a protein region from residue  
9 627 to 1092, was amplified using the primers CDC7forw2 (5'-GAATTCTGGATCCGTGGGTTTCAGCC-3', the  
10 EcoRI site is underlined) and CDC7rev2 (5'-GCGGCCGCTTAGTCCATGGAGGTGAAGTATGG-3', the NotI site is  
11 underlined). PCR reaction were performed in a final volume of 50  $\mu$ l using 50  $\mu$ l of 10X buffer (Stratagene,  
12 La Jolla, CA, USA), 50  $\mu$ M dNTPs (Takara), 20 pmols of each primer and 2.5 units of Pfu Ultra High Fidelity  
13 (Stratagene). Reactions were performed on a T-Personal Thermocycler (Biometra Corporation).  
14 Amplification conditions were: one cycle at 95°C for 5 min; 30 cycles at 95°C for 30 sec, 55°C for 30 sec and  
15 72°C for 2 min; and one cycle at 72°C for 7 min. For the *in vivo* expression in *G. duodenalis* parasites, the full  
16 length *cdc7* PCR fragment was cloned in the BamHI/PspOMI-digested pTUB-FLAGpac vector. For the  
17 expression in *E. coli* of GST-fused recombinant protein, the CDC7-II fragment was cloned in the EcoRI/NotI  
18 digested pGEX-6P1 vector (GE Healthcare, Little Chalfont, England). The plasmids p14-X for the bacterial  
19 expression of GST-fused g14-3-3 (GST-g14-3-3) was described previously.<sup>23</sup>

#### 2.4 RNA reverse transcription and Real-time quantitative PCR

20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
Condition for cDNA synthesis and Real-Time PCR were as described.<sup>25</sup> The final mRNA levels of the  
studied genes were normalized to *gap1* (glyceraldehyde 3-phosphate dehydrogenase) gene expression using  
the relative quantification method (Roche Diagnostics). Graphed values were estimated from five  
independent experiments each repeated in triplicate. The *cwp1*, *gap1* and *gcdc7* genes were amplified,  
respectively, using the primers RT-CWP1forw (5'-CTGCATCAATGAGCTTCAATT-3') and RT-CWP1rev (5'-  
TGCTGACAGCTGATTGC-3'), RT-GAP1forw (5'-ACAGGTCGCTTTACAACGAAG-3') and RT-GAP1rev (5'-  
AGATGATGACACGCTTGACAG-3'), RT-CDC7forw (5'-CTCAGGATCTGCCAGAAGGA-3') and RT-CDC7rev (5'-  
GAGAGAACGGTTCGTCGATATT-3').



1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

### 2.5 Expression and purification of the recombinant proteins

*E. coli*-transformed cells were grown in SOB medium at  $OD_{600} = 0.6-0.8$ , and the expression of recombinant proteins was induced in the presence of 0.5 mM IPTG (isopropyl-thio- $\beta$ -D-galactoside) at 37°C for 4 h. All GST-fused proteins were purified by affinity chromatography on glutathione-sepharose 4B (GE Healthcare) and eluted with 10 mM reduced glutathione (pH 8.0).

### 2.6 Production of Polyclonal Antibodies

Two BALB/c mice (Charles River Laboratories International, Inc.; MA, USA) were immunized intraperitoneally on days 0, 21, and 42 with 50  $\mu$ g of GST-CDC7-II fusion protein resuspended in 150  $\mu$ l of phosphate-buffered saline (PBS) and emulsified with an equal volume of Freund's complete adjuvant (Sigma Aldrich, St. Louis, Missouri, USA) (at day 0) or Freund's incomplete adjuvant (Sigma Aldrich) (at day 21) or without any adjuvant (at day 42). Blood was collected prior to initial immunization and after each boost from the tail vein, the serum fraction was assayed for specific antibody content.

### 2.7 Preparation of Giardia proteins

Total soluble proteins were prepared according to<sup>23</sup> with modifications. Briefly,  $2 \times 10^9$  trophozoites or encysting parasites were collected by chilling on ice and were washed three times with cold PBS and the cell pellet was frozen at -70°C overnight. Cells were resuspended in two volumes of extraction buffer (30 mM Tris-HCl, 1 mM DTT, and 1 mM EDTA, pH 7.4), supplemented with protease-inhibitor cocktail (P8340, Sigma-Aldrich) and phosphatase-inhibitor cocktail (P2850, Sigma-Aldrich) and then destroyed by sonication. The lysate was centrifuged at 24,000 g for 30 min at 4°C, and the supernatant was collected and designed as "soluble" fraction (S). The sediment containing the membranous material was washed twice with cold extraction buffer and centrifuged both times for 30 min at 24,000 g at 4°C. The pellet was then resuspended in 1 volume of extraction buffer supplemented with 2% octylglucoside and constantly stirred at 4°C overnight. Further solubilization was achieved by sonication (5 times for 30 sec at 50% power and 20% duty cycle) and centrifugation was performed at 24,000 g for 30 min at 4°C. The supernatant was collected and designed as "membrane" fraction (M). The protein concentration was measured with the method of Bradford (Pierce, Rockford, IL, USA) and the material was stored at -70°C.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60**2.8 Western blot analysis**

Proteins were separated on SDS-PAGE and transferred onto nitrocellulose membrane with 39 mM glycine, 48 mM Tris, 0.1% SDS, and 10% methanol, using a semidry apparatus (BioRad, Hercules, CA, USA). Membranes were blocked with 5% skin milk in T-TBS (20 mM Tris-HCl, pH 7.5, 100 mM NaCl, 0.05% Tween 20) for 1h and then incubated with the primary antibody (Ab) in T-TBS/2% skin milk. After incubation with an appropriate HRP-conjugated secondary Ab (1:2000) the interaction was revealed by chemiluminescence (Millipore, Billerica, MA, USA). The mouse M2 anti-FLAG mAb (Sigma-Aldrich), the rabbit anti- $\alpha$ TUB mAb (AB-11661, Immunological Sciences, Italy) were used at a dilution of 1:500; the mouse anti-CWP mAb (1E10)<sup>27</sup> was used 1:20; the rabbit N14 ( $\alpha$ -g14-3-3) antiserum<sup>23</sup> was used at a 1:5000 dilution; the mouse anti-gCDC7 polyclonal serum, the mouse anti-gCLH (chlatriin heavy chain),<sup>28</sup> the mouse anti-g $\beta$ COP (coatamer beta)<sup>28</sup> and the rabbit anti-gPGM (phosphoacetylglucosamine mutase)<sup>29</sup> were used at a dilution of 1:1000.

**2.9 Affinity purification**

In two independent experiments, FLAG-fusion proteins were purified using mouse anti-FLAG M2 mAb covalently bound to agarose beads (Sigma-Aldrich). The column was prepared following the manufacturer's instructions. Briefly, the resin was activated with 0.1 M glycine pH 3.5, washed twice with five volumes of PBS buffer and equilibrated in KH-Tween buffer (20 mM Hepes-KOH, pH 7.6, 225 mM KCl, 5 mM MgCl<sub>2</sub>, 0.1 mM EDTA, 0.04% Tween 20). *Giardia* soluble protein fractions, from trophozoite (10 mg) or encysting parasites (5 mg) were incubated with anti-FLAG beads at 4°C for 3 h and washed with 100 resin bed volumes of KH-Tween buffer. Finally, FLAG-fusion proteins were eluted from the resin by incubation with 200  $\mu$ M synthetic FLAG-peptide (Sigma-Aldrich) at 4°C for 1 h. The collected materials were stored at -70°C until use.

**2.10 Mass spectrometry analysis**

Aliquots of FLAG-g14-3-3 or FLAG-gCDC7 complexes and controls, affinity purified as previously described, were separated on a 1D-gel NuPAGE 4-12% (Novex, Invitrogen, Carlsbad, CA, USA) run in MOPS buffer and stained with the Colloidal Blue Staining kit (Invitrogen). The whole lanes were cut in 24



1  
2  
3 homogeneous slices (for the FLAG-gCDC7 complexes only selected slices were analyzed) and processed for  
4  
5 cysteine residues reduction and alkylation by iodoacetamide (Sigma-Aldrich) and trypsin (Promega  
6  
7 Corporation, Madison, WI, USA) digestion over night, as elsewhere described.<sup>30</sup> Peptide mixtures were  
8  
9 analyzed by nanoflow reversed-phase liquid chromatography tandem mass spectrometry (RP-LC-MS/MS)  
10  
11 using an HPLC Ultimate 3000 (DIONEX, Sunnyvale, CA U.S.A) connected on line with a linear Ion Trap (LTQ-  
12  
13 XL, ThermoElectron, San Jose, CA). Peptides have been desalted in a trap column (Acclaim PepMap 100  
14  
15 C18, LC Packings, DIONEX) and then separated in a reverse phase column, a 10 cm long fused silica capillary  
16  
17 (Silica Tips FS 360-75-8, New Objective, Woburn, MA, USA) slurry-packed in-house with 5  $\mu\text{m}$ , 200  $\text{\AA}$  pore  
18  
19 size C18 resin (Michrom BioResources, CA). Peptides were eluted using a linear gradient from 96% A ( $\text{H}_2\text{O}$   
20  
21 with 5% acetonitrile and 0.1% formic acid) to 60% B (acetonitrile with 5%  $\text{H}_2\text{O}$  and 0.1% formic acid) for 40  
22  
23 min, at 300nl/min flow rate. Analyses were performed in positive ion mode and the HV Potential was set up  
24  
25 around 1.7-1.8kV. Full MS spectra ranging from m/z 400 to 2000 Da were acquired in the mass  
26  
27 spectrometer operating in a data-dependent mode in which each full MS scan was followed by five MS/MS  
28  
29 scans where the five most abundant molecular ions were dynamically selected and fragmented by CID  
30  
31 using a normalized collision energy of 35%. Target ions already fragmented were dynamically excluded for  
32  
33 30s. Tandem mass spectra were matched against *Giardia* protein database (Giardia DB version 1.2)  
34  
35 downloaded from the web site <http://www.giardiadb.org/giardiadb> and through SEQUEST algorithm<sup>31</sup>  
36  
37 incorporated in Bioworks software (version 3.3, Thermo Electron) using fully tryptic cleavage constraints  
38  
39 with the possibility to have one miss cleavage permitted, static carbamidomethylation on cysteine and  
40  
41 methionine oxidation as variable modification. Data were searched with 1.5 Da and 1 Da tolerance  
42  
43 respectively for precursor and fragment ions. A peptide has been considered legitimately identified when it  
44  
45 achieved cross correlation scores of 1.8, 2.5 and 3 respectively for charge states 1,2 and 3, and a peptide  
46  
47 probability cut-off for randomized identification  $p < 0.001$ . The single-peptide based protein identifications  
48  
49 for each replica experiments (1 and 2) for trophozoite sample (0h) and 12h encysting parasites (12h) are in  
50  
51 separate lists in Supplemental Data. Protein and peptide false discovery rate (FDR) has been calculated  
52  
53 dividing the number of false hits by the number of positive hits where the false hits are evaluated using a  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 decoy database directly constructed by the Bioworks software on the same database used for the target  
4  
5 search and adopting the same scoring criteria. EmPAI (Exponentially modified Protein Abundance Index)  
6  
7 values have been calculated for each protein using the following formula:  $emPAI = 10^{PAI} - 1$ , where PAI  
8  
9 is calculated dividing the number of the peptides detected during MS analysis normalized for the  
10  
11 theoretical number of detectable peptides<sup>32</sup> calculated assuming the possibility to measure only tryptic  
12  
13 peptides containing at least 5 aminoacids and no more than one miss-cleavage.  
14

#### 15 16 17 **2.11 Protein sequence analysis and statistical significance assessment**

18  
19 Bioinformatic and statistical methods are detailed in Supplemental data.  
20

#### 21 22 **2.12 Overlay assay**

23  
24 The overlay assay was carried out according to Lalle et al., 2006.<sup>23</sup> The proteins were separated,  
25  
26 transferred and blocked as described. The membrane was then incubated at 4°C overnight in HT buffer  
27  
28 with 2% ECL-blocking agent (GE Healthcare) and with 5 µg/ml of GST-g14-3-3. After washing with HT buffer,  
29  
30 the membrane was first incubated with the goat HRP-conjugated anti-GST mAb (GE Healthcare) at a 1:2000  
31  
32 dilution. The interaction was revealed with the ECL system (GE Healthcare).  
33

#### 34 35 **2.13 Blue Native PAGE (BN-PAGE)**

36  
37 3-12% BN-PAGE (Invitrogen) was carried out with anti-FLAG immunopurified material from trophozoites  
38  
39 and encysting parasites of the WB-C6 parental strain, FLAG-g14-3-3 and FLAG-gCDC7 transgenic lines  
40  
41 according to the manufacturer. This technique allows to separate very high molecular weight multiprotein  
42  
43 complexes. Gel were run in Running buffer (0.002% coomassie G-250, 50mM BisTris/50mM Tricine, pH 6.8)  
44  
45 at 150V for approximately 2h and stained with Silver Staining kit (Invitrogen). Gel bands were excised and  
46  
47 treated for mass spectrometry analysis as described in paragraph 2.10.  
48  
49

#### 50 51 **2.14 Confocal Laser Scanning microscopy**

52  
53 Trophozoites or encysting cells were prepared as previously described.<sup>23</sup> Polyclonal rabbit N14  
54  
55 antiserum (anti-g14-3-3), polyclonal mouse anti-gCDC7, mouse Cy3-conjugated anti-FLAG mAb (Sigma-  
56  
57 Aldrich) and mouse FITC-conjugated anti-CWP mAb (Waterborne Inc.; New Orleans, LA, USA) were used at  
58  
59 dilutions of 1:20, and Alexa-Fluor 594-conjugated anti-rabbit secondary Ab and Alexa-Fluor 546-conjugated  
60

1  
2  
3 anti-mouse secondary Ab (Invitrogen) were used at a 1:500 dilution. After staining, coverslips were  
4 extensively rinsed and then mounted on the microscope slide by using Vectashield<sup>®</sup> mounting medium  
5  
6 (Vector Laboratories Inc.; Burlingame, CA, USA) containing 300 nM of 4',6-diamidino-2-phenylindole (DAPI)  
7  
8 before Confocal Laser Scanning Microscopy (CLSM) analyses. The observations were performed on a Leica  
9  
10 TCS SP2 AOBS apparatus, utilizing excitation spectral laser lines at 405, 488 and 594 nm, properly tuned by  
11  
12 acousto-optical tunable filter (AOTF). The emission wavelengths were selected by a proper setting of the  
13  
14 spectral detection system. Image acquisition and processing were conducted by using the Leica Confocal  
15  
16 Software (Leica Lasertechnik GmbH, Heidelberg, Germany). Signals from different fluorescent probes were  
17  
18 taken in sequential scan mode, and co-localization was detected in yellow. The image processing was  
19  
20 performed using the Huygens software (Scientific Volume Imaging BV, Hilversum, The Netherlands).  
21  
22 Different fields of view (>200 cells) were analyzed on the microscope for each labeling condition and  
23  
24 representative results are shown.

### 30 3. Results

#### 31 3.1 Isolation and Proteomic Identification of *g14-3-3-associated Ligands*

32  
33 In order to achieve a physiological expression of the FLAG-tagged *g14-3-3*, as comparable as  
34  
35 possible to the endogenous protein, a 100 bp region upstream the *g14-3-3* starting codon was cloned at  
36  
37 the 5' of the FLAG-*g14-3-3* coding sequence instead of the of the *gdh* promoter in the *pya*-FLAG14-3-3  
38  
39 vector.<sup>24</sup> As described for other *Giardia* promoters,<sup>33</sup> this non-coding region, contains a putative AT-rich  
40  
41 sequence, generally corresponding to the transcription initiation site, and three putative g-CAB elements  
42  
43 (*Giardia* CAT box) upstream from the AT-rich sequence (Fig. 1A). As shown by immunoblotting with anti-  
44  
45 FLAG mAb, the FLAG-*g14-3-3* was efficiently expressed during the vegetative (trophozoite, T) and  
46  
47 encystation (12h) stages (Fig. 1B, panel anti-FLAG) of *Giardia* parasites transfected with the *pya*-  
48  
49 P14\_FLAG14-3-3, thus indicating that the 100 bp upstream region encompassed a true *Giardia* promoter  
50  
51 sufficient to drive the transcription, and then the expression, of the FLAG-*g14-3-3*. The amount of the  
52  
53 exogenous protein was anyhow lower than that of the endogenous one (Fig. 1B, panel anti-*g14-3-3*).  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 Notably, the intracellular localization of the FLAG-g14-3-3 (data not shown) was almost identical to that  
4  
5 previously reported for the FLAG-g14-3-3 expressed under the *gdh* promoter.<sup>24</sup>  
6

7  
8 FIG. 1  
9

10 To obtain a picture of the interaction networks established by g14-3-3 in different *Giardia* life  
11 stages, the FLAG-tagged protein was then immunoprecipitated (IP) using anti-FLAG mAb, from *Giardia*  
12 trophozoites and parasites collected at 12h post encystation induction. At the latter stage, a higher fraction  
13 of g14-3-3 is underglycosylated, which impinges on its intracellular localization.<sup>23</sup> The IP material from the  
14 FLAG-g14-3-3 transfected line showed a complex pattern of protein bands, more than the control WB-C6 IP,  
15 and g14-3-3 protein visually dominate the PAGE display as confirmed by WB with anti-g14-3-3 Ab (Fig. 1C,  
16 panels coomassie and g14-3-3, respectively). The overlay assay, performed with the recombinant GST-g14-  
17 3-3, indicated that the IP proteins from the transgenic line, but not from the control line WB-C6, were  
18 enriched in g14-3-3 binding targets (Fig. 1C, panel overlay). The specificity of the overlay assay was  
19 demonstrated by the suppressive effect of the competing A8Ap phosphopeptide (Fig. 1C, panel  
20 overlay+A8Ap), that reproduces a mode-1 14-3-3 binding motif.<sup>13</sup>  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34

35 To identify the putative g14-3-3 binding proteins, immunoprecipitates from FLAG-g14-3-3-  
36 transfected and control lines were resolved on polyacrylamide gels and whole lanes were analyzed by LC-  
37 MS/MS. Both the immunoprecipitation and mass spectrometry analyses were repeated in two entirely  
38 independent experiments. Proteins identified only in the FLAG-g14-3-3 precipitate are proposed as putative  
39 g14-3-3 interactors. In general, proteins detected in both the FLAG-g14-3-3 and control lanes should be  
40 considered common, non-specific contaminants. Hence, their abundance in the transfected and control  
41 precipitates should be similar and consequently the number of peptides assigned to the same protein,  
42 reflected in the emPAI value, should be comparable.<sup>32</sup> We analyzed the emPAI values (see Experimental  
43 Procedures) of all 110 presumptive contaminants and calculated, for each protein, the ratio between the  
44 emPAI in the FLAG-g14-3-3 precipitate and in its control at 0 or 12h after encystation. We found that these  
45 values were normally distributed and centered on an average ratio of  $1.2 \pm 0.9$  S.D. Thus, we regarded  
46 presumptive contaminants with an emPAI ratio  $> 3$  (mean value plus twice the standard deviation) as  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 putative FLAG-g14-3-3 interactors. Thus, we included in the list of candidate 14-3-3 interactors the proteins  
4  
5 that, though present also in the control precipitate, had an emPAI ratio > 3.  
6

7  
8 Altogether, as shown in Supplemental Table 1, 296 proteins were identified in the IP from FLAG-  
9  
10 g14-3-3 trophozoites, with 162 identifications (54%) obtained with more than one peptide. Fifty proteins  
11  
12 were identified in the IP from FLAG-g14-3-3 parasites at 12h of encystation, with 13 out of 50 (27%)  
13  
14 assigned with more than one peptide. 33 proteins were common to the two IP sub-sets. Total peptide and  
15  
16 protein FDRs were 1% and 4%, respectively. The protein FDR was determined exclusively by the inclusion of  
17  
18 proteins identified with only one peptide, since in the decoy database no identifications were obtained  
19  
20 with more than one peptide.  
21  
22

### 23 **3.2 In silico detection and analysis of 14-3-3-binding motifs**

24  
25 Since the interaction of proteins target with 14-3-3 occurs, in most of the cases, through well  
26  
27 defined phosphorylated motifs,<sup>2</sup> the identified proteins were scanned by pattern matching for the presence  
28  
29 of at least one of 22 14-3-3 binding motifs (Supplemental Table 2) collected from different databases (see  
30  
31 Supplemental Materials and Methods). As resulted, a subset of 14 motifs were detected and the majority of  
32  
33 the identified proteins contained at least one of these 14-3-3 binding motifs (Supplemental Table 1), thus  
34  
35 supporting the overlay assays (Fig. 1C).  
36  
37

38  
39 To verify if any of the detected motifs was over-represented in the list of the putative 14-3-3  
40  
41 targets in comparison to the whole *Giardia* genome we have calculated the Hypergeometric p-value for  
42  
43 each of the motifs. Four of the 14 binding motifs possess a hypergeometric p-value below 0.01  
44  
45 (Supplemental Table 3) and are over-represented in list of the putative 14-3-3 targets (Supplemental Fig.  
46  
47 S1A), thus further confirming the existence of true positives in the pool of proteins co-purified with the  
48  
49 FLAG-g14-3-3 and suggesting direct interactions. However, the 6% of the identified proteins (19 out of 313)  
50  
51 excluding the g14-3-3 itself, did not harbor any canonical 14-3-3 binding motif (indicated with an asterisk  
52  
53 in Supplemental Table 1), thus suggesting that this pool of proteins could indirectly associate with FLAG-  
54  
55 g14-3-3 via other g14-3-3-associated proteins. Mostly of this 19 proteins were identified with a single  
56  
57 peptide and/or classified as part of the ribosomal multiprotein complex. However, non-canonical 14-3-3  
58  
59  
60



1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

binding motifs could be also present in some of these proteins. Nevertheless, the presence of a 14-3-3 binding motif is not sufficient to ensure that a protein interacts with 14-3-3 through it. In fact, two of the identified proteins, namely gTLL3 and gDIP2, even possessing a putative 14-3-3 binding motif are not 14-3-3 binding targets but act as polyglycylase and deglycylase of the g14-3-3, respectively.<sup>25</sup>

We further analyzed the core motif (R..S), including five upstream and downstream residues, in the experimentally identified proteins co-purified with g14-3-3 (BMC in Supplemental Fig. S1A) with those in the *Giardia* genome that posses the motif but were not identified in our experiments (MC minus BMC in Supplemental Fig. S1A). A Weblogo plot showed that in the experimentally identified proteins, hydrophobic residues (L, I, M and V) surrounding the core motif R..S were favored (positions -4, -2, -1 and +2) as well as charged residues (D) in the rather distant positions -8 and +3 (Supplemental Fig. S1B). Respect to the negative sample, cysteines (C) are not found, which might indicate that flexibility is required for the motif to mediate the interaction. Our result was quite distant from the frequency plot of more than 200 reported 14-3-3 binding site of plant and animal, where the most common residues were Leu and Arg at position -5 and Ser and Pro at position +2.<sup>20</sup> This discrepancy could mainly be the result of the comparison of all R..S containing sequences in our analysis rather than only sequences experimentally proved to mediate 14-3-3 binding.<sup>20</sup>

Since it has been observed that more than 90% of characterized 14-3-3 protein partners contain disordered regions and almost all 14-3-3-binding sites are inside disordered regions<sup>20,34</sup> we evaluated, using IUPred, the presence of structured and disordered regions in the proteins identified. Intriguingly, almost 25% of the identified proteins having a (R..S) core motif were predicted to be disordered (a motif is considered as disordered if all its fixed residues have a disorder score > 0.5) and 82% of the them were predicted also to be exposed. We further try to asses if the putative 14-3-3 binding motifs detected in the identified proteins were exposed or not on the basis of the protein tertiary structure. When the structure of a protein is known, solvent accessibility can be directly computed from its atomic coordinates. Only few of the giardial proteins identified in our list have a known structure (i.e. the fructose-bisphosphate aldolase).<sup>35</sup> For the remaining ones, one could either predict the accessibility from the sequence alone or, if

1  
2  
3 a sufficiently close homologue of known structure is available, build a comparative model of the structure  
4  
5 of the proteins and compute approximate accessibility values from them. We were able to model 87  
6  
7 proteins out of 314 (data not shown) and the reliability of the obtained models were calculated with  
8  
9 QMEAN. A 57% of the models displayed a QMEAN score >0.6 (QMEAN score range from 0 to 1). Based on  
10  
11 such model, 84 out of 175 motifs present in the modeled proteins were predicted to be exposed to solvent  
12  
13 with an average accessibility value above 25% (data not shown). As an example, the models with the  
14  
15 localization of the putative 14-3-3 binding motifs (Supplemental Table 1) of V-ATP synthase  $\beta$   
16  
17 (GL50803\_12216) and of malic enzyme (GL50803\_14285) are reported (Supplemental Fig. S2).  
18  
19  
20

### 21 **3.3 Functional classification of 14-3-3-associated proteins**

22  
23  
24 Even if, at the time of our analysis, the GO terms annotation for *Giardia* was not deep enough  
25  
26 (Supplemental Table 1) all the identified proteins have been gathered into general categories of function  
27  
28 and related processes (Table 1) combining the annotation at Giardia DB 1.2 and BLAST homology search.  
29  
30 Nevertheless, an high percentage of the proteins identified in both trophozoite and 12h encysting parasite  
31  
32 subsets (24% and 17.5%, respectively) were classified as protein of unknown function. In this percentage  
33  
34 where also included the uncharacterized protein 21.1, a *Giardia*-specific gene family of ankyrin-repeat  
35  
36 domain containing proteins.<sup>22</sup> Proteins classified in the *Giardia* database with unknown function but sharing  
37  
38 homology with characterized proteins or contained functional protein domains were however assigned to  
39  
40 functional groups. As listed in Table 1 (see references column) many proteins with similar name or function  
41  
42 have been described as 14-3-3-interacting proteins in other organisms, indicating that the protein-g14-3-3  
43  
44 complexes herein identified are biologically significant. As examples, plant and animal 14-3-3s bind and  
45  
46 negatively regulate functional homologs of the ATP synthase  $\beta$ ,<sup>13,17,48</sup> and  $\alpha$ - and  $\beta$ -tubulins have been  
47  
48 identified in complex with 14-3-3 in multiple *in vivo* and *in vitro* interaction screens.<sup>13,17,19,39</sup> The largest part  
49  
50 of proteins identified in *Giardia* trophozoite sample were not found in the 12h encysting parasite sample,  
51  
52 suggesting that many of the observed interactions take place in a stage-dependent fashion. Indeed, few  
53  
54 proteins were exclusively identified in the encystation sample (Table 1 and Supplemental Table 1).  
55  
56  
57  
58  
59  
60 Nonetheless, the percentage distribution of g14-3-3 interactors into general function/processes categories



1  
2  
3 was comparable between the two stages examined (Supplemental Fig. S3), with the exception of metabolic  
4  
5 and energy-related processes, which were relatively more represented during encystation.  
6

7  
8 Table 1  
9

### 10 **3.4 Confirmation of some 14-3-3 Interactions**

11  
12 To confirm some of the identifications achieved by mass spectrometry analysis, IP material was  
13  
14 assayed by western blotting, whenever suitable antibodies were available. *Giardia*-specific antibodies  
15  
16 against the clathrin heavy chain (gCLH), the coatamer protein complex I subunit beta ( $\beta$ COP) and the  
17  
18 phosphoacetylglucosamine mutase (gPGM) were used. Moreover, a central region of the putative gCDC7  
19  
20 homolog protein (residue 627-1092) was expressed as a GST-fusion and a mouse polyclonal serum was  
21  
22 raised against the GST-released recombinant protein. As shown in Fig. 2, and in agreement with MS data,  
23  
24 bands corresponding to the molecular size of gCDC7 and gCLH were visible in all IP samples from FLAG-g14-  
25  
26 3-3 transfectants. As well, a band compatible with the g $\beta$ COP in the throphozoite sample (T) and a one  
27  
28 compatible gPGM in the 12h encysting parasite sample (12h) were immunodecorated in the IP materials  
29  
30 from FLAG-g14-3-3 transgenic line. No signal with any of the antibodies was observed in the control IPs  
31  
32 from WB-C6 parental strain.  
33  
34  
35  
36  
37

38 Fig. 2  
39

### 40 **3.5 Expression and localization of the putative gCDC7 protein**

41  
42 Among the identified proteins, the putative gCDC7 homolog, a predicted protein of 1697  
43  
44 aminoacids, was one of the most abundant proteins in all the identification experiments, with a high  
45  
46 sequence coverage (up to 32%). In eukaryotes, the serine/threonine protein kinase Cdc7 associates with  
47  
48 the regulatory subunit Dbf4 (Dumbbell-forming 4)<sup>55</sup> to form the so called DDK (Dbf4-Dependent Kinase).  
49  
50 The activity of DDK, in combination with cyclin-dependent kinases, has been demonstrated to be essential  
51  
52 for cell cycle progression and in response to DNA damage or replication fork stalling.<sup>56</sup> Moreover, Cdc7 has  
53  
54 been recognized has a novel target for cancer therapy<sup>57</sup> and proposed as a potential target for anti-giardial  
55  
56 drugs.<sup>22</sup> Due to the potential relevance for the parasite biology, the gCDC7 protein and its interactions with  
57  
58 g14-3-3 were further studied. As shown in Supplemental Fig. S4, the N-terminal portion of the gCDC7,  
59  
60

1  
2  
3 encompassing the first 450 residues, contains a protein kinase domain that, as inferred by BLAST analysis  
4  
5 (data not shown), shares a 65-70% sequence homology with the kinase domain of Cdc7 proteins of other  
6  
7 organisms.<sup>55</sup> Three amino acids segments of undefined function, that interrupt the kinase domain at  
8  
9 conserved locations,<sup>56</sup> were also present in the gCDC7 as in other Cdc7 homologs (Supplemental Fig. S4). In  
10  
11 contrast, no other conserved domains are present in the C-terminal half of the protein. Several potential  
12  
13 14-3-3 binding motifs, mostly outside of the kinase domain, were identified and one of them was located  
14  
15 inside a putative Nuclear Localization Signal (NLS, Supplemental Fig. S4).  
16  
17

18  
19 Fig. 3

20  
21 The gene expression profile of *gcdc7* was investigated in the parental WB-C6 strain at different time  
22  
23 points after the induction of encystation (Fig. 3A). The relative gene expression level of *gcdc7*, normalized  
24  
25 against the constitutive *gap1* (glyceraldehyde 3-phosphate dehydrogenase) gene was comparatively  
26  
27 constant during encystation, except for a decrease around 3h post encystation induction. As expected, the  
28  
29 expression of the encystation-induced *cwp1* (cyst wall protein 1) gene was strongly upregulated, starting  
30  
31 from 3h post-encystation induction. The expression and localization of the gCDC7 protein was then studied  
32  
33 by western blotting, both in trophozoites and in encysting parasites. A 190 kDa band was detected in the  
34  
35 analyzed parasite stages (Fig. 3B), moreover, a decrease in the gCDC7 amount was observed at 3h and at  
36  
37 24h post encystation induction (Fig. 3B). The gCDC7 was then localized in trophozoites, encysting parasites  
38  
39 and cysts. The production of encysting specific vesicles (ESVs) was used as easily observable marker of  
40  
41 *Giardia* encystation and anti-cyst wall protein (CWP) mAb was used to stain ESVs and cyst wall. In fixed and  
42  
43 permeabilized trophozoites (Fig. 3C) the anti-gCDC7 serum produced a spotted labeling within the parasite  
44  
45 cell body with a strong signal corresponding to the axonemal exit points, or flagellar pores, as suggested by  
46  
47 staining of the eight flagella with anti- $\alpha$ -Tubulin (Fig. 3C, panel a). A partial co-localization with g14-3-3  
48  
49 was more evident in the cell body (yellow color) than at the flagellar pores (Fig. 3B, panel b). No significant  
50  
51 labeling of the nuclei or other recognizable structures could be reported. In encysting parasites (Fig. 3C,  
52  
53 panels c and d) the labeling of flagellar pores with anti-gCDC7 serum progressively disappeared (it is  
54  
55 completely absent in panel d), whereas a spotted staining was visible inside of the nuclei. A partial co-  
56  
57  
58  
59  
60

1  
2  
3 localization with g14-3-3 was also evident inside the nuclei and in the cell body (Fig. 3C, panel d). The  
4  
5 nuclear localization of the gCDC7 is consistent with the role in DNA replication reported for other CDC7  
6  
7 protein kinases. In the late stage of encystation both nuclei divide and the DNA is replicated during a  
8  
9 modified cell cycle, involving karyokinesis without cytokinesis or endoreplication.<sup>59,60</sup> In cysts (Fig. 3C,  
10  
11 panels e and f) the gCDC7 localized preferentially at the cyst periphery, below the cyst wall, where the  
12  
13 signal partially overlapped with that of g14-3-3. The association of gCDC7 with the membrane fraction was  
14  
15 also confirmed by immunoblotting. Anti-gCDC7 immunodecorated a band in the insoluble protein fraction  
16  
17 (membranous material) from 12h encysting parasite and in the soluble protein fraction at both stages (Fig.  
18  
19 3D). As previously demonstrated, g14-3-3 was present in soluble and insoluble fractions in all samples (Fig.  
20  
21 3D), (Lalle et al., 2006). Stage specificity was confirmed by the expression of CWP in encysting parasites.  
22  
23

24  
25  
26 Fig. 4  
27

### 28 **3.6 Characterization of the g14-3-3-gCDC7 complex and identification of the putative gDBF4 homolog**

29  
30 To confirm the interaction between g14-3-3 and gCDC7 we generated a *Giardia* transgenic line  
31  
32 expressing an N-terminally FLAG-tagged gCDC7 (FLAG-gCDC7) under the  $\alpha$ -tubulin constitutive promoter.  
33  
34 Expression of the tagged protein was confirmed by western blotting (Fig. 4A). As shown by anti-gCDC7 and  
35  
36 anti-g14-3-3 immunoblotting (Fig. 4B), anti-FLAG mAb immunoprecipitated the FLAG-CDC7 from  
37  
38 transfected parasite extracts together with the g14-3-3. To prove the direct interaction between g14-3-3  
39  
40 and gCDC7, overlay experiments were performed on the IP materials from trophozoite using the GST-g14-  
41  
42 3-3 as bait. A band doublet in the range of 160-260 kDa was visible only in the IP material from the FLAG-  
43  
44 gCDC7 transgenic line, with the lower band compatible with the molecular size of FLAG-gCDC7 (Fig. 4C). To  
45  
46 identify the two protein bands recognized by the GST-g14-3-3, the protein mixture present in the IP  
47  
48 material in the portion of the gel between 160 and 260 kDa were analyzed by LC-MS/MS. As reported in  
49  
50 Supplemental Table 4, two proteins were exclusively enriched in the FLAG-gCDC7 IP material: FLAG-gCDC7  
51  
52 itself and a protein of 2193 amino acids (GL50803\_94117), with a predicted MW of 241 kDa and unknown  
53  
54 function. Remarkably, the GL50803\_94117 protein (Table 1) co-purified also with FLAG-g14-3-3 in both  
55  
56 trophozoite and 12h encysting parasite subsets. Protein motif scanning identified, in the central region of  
57  
58  
59  
60

1  
2  
3 the protein (residues 976-1036), a DBF4-type zinc finger (Pfam PF07535) or motif C, generally present in the  
4  
5 Cdc7-regulatory subunit Dbf4 and necessary for the interaction with the Cdc7 protein.<sup>61,62</sup> Comparison of  
6  
7 the amino acid sequences of the GL50803\_94117 protein, here termed gDbf4, with Cdc7 regulatory  
8  
9 subunits from various eukaryotes (Supplemental Fig. S5) revealed also the presence of the so called Dbf4  
10  
11 motifs N and M (Ogino et al.; 2001), thus supporting a possible role of the protein as gCDC7-regulatory  
12  
13 subunit. However, compared with other Dbf4 homologs, the putative gDBF4 possesses an extended C-  
14  
15 terminal portion without known conserved domains.  
16  
17

18  
19 To prove the presence of gCDC7, gDbf4 and g14-3-3 in the same complex immunoprecipitated  
20  
21 material from FLAG-gCDC7- and FLAG-g14-3-3-transfected trophozoites were run on native PAGE, a  
22  
23 technique allowing the separation of intact protein complexes on polyacrylamide gel, and protein  
24  
25 complexes were then analyzed by mass spectrometry. As shown in Fig. 4D, a band around 650-700 kDa was  
26  
27 coomassie-stained only in the FLAG-gCDC7 and FLAG-g14-3-3 IP material, but not in the control WB-C6 IP.  
28  
29 MS analysis revealed the simultaneous presence in this band of gCDC7, gDbf4 and g14-3-3 that were  
30  
31 recognized (data not shown).  
32  
33  
34

#### 35 36 37 38 **4. Discussion**

39  
40 Here, we report the first proteomic analysis of the multiprotein complexes established by the single  
41  
42 14-3-3 isoform in the phylogenetically basal eukaryote *G. duodenalis*. We have previously suggested that  
43  
44 g14-3-3 is part of a complex network of *in vivo* interactions, on the basis of the large number of proteins it  
45  
46 binds *in vitro*.<sup>23</sup> In the present work, an immunoaffinity approach has been used to capture the complexes  
47  
48 formed *in vivo* by a FLAG-tagged version of g14-3-3, ectopically expressed under its own promoter region.  
49  
50 We have identified more than 300 proteins co-purifying with g14-3-3 highly enriched in proteins containing  
51  
52 one or multiple putative 14-3-3-binding motifs, some of which are predicted to localize in regions suitable  
53  
54 for g14-3-3 binding. However, it is likely that only a fraction of the identified proteins are *bona fide* direct  
55  
56 14-3-3 interactors, whereas the others bind g14-3-3 indirectly. In fact, subunits of well characterized  
57  
58  
59  
60

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

multiprotein complexes (e.g. the small and large ribosomal subunits) are present among the g14-3-3-co-precipitated molecules.

Our data, combined with the wide 14-3-3 literature, highlight that certain 14-3-3 interactions occur in *Giardia* as well as other eukaryotes, thus confirming that the involvement of 14-3-3 in defined biological processes has been well conserved throughout eukaryotic evolution. However, some well-known 14-3-3 interactors, such as histone acetyltransferases (HACs) and histone deacetylases (HDACs)<sup>12,19</sup> were not identified here. In the case of HDACs, 14-3-3 interaction has been proven to occur only with animal class IIa deacetylases<sup>63</sup> and plant HD2 HDAC family,<sup>19</sup> whereas *Giardia* harbors a reduced repertoire including only one class I HDAC and about five Sirtuin family members.<sup>64</sup>

In agreement with our previous observations showing that g14-3-3 point mutations cause remarkable alterations of encystation,<sup>23,24</sup> the information collected in this work links g14-3-3 to processes that are central to the survival of the parasite and are deeply modified during *Giardia* encystation, as discussed here below. In *Giardia* trophozoites, the involvement of g14-3-3 in protein trafficking and vesicular transport is suggested by its co-precipitation along with subunits of the COPI and COPII complexes, that function in ER-to-Golgi transport and with the adaptor protein complex AP2, involved in endocytosis.<sup>65</sup> Furthermore, g14-3-3 coprecipitates with RABs, such as Rab11, belonging to the Ras superfamily of small GTPases, and guanine-nucleotide exchange factors (GEFs). In yeast and higher eukaryotes, binding of 14-3-3 to cargo proteins is required for the ER-to-Golgi transport of a number of multimeric membrane proteins. 14-3-3s act as sensors of the correct assembly of multimers and lead to proper export to plasma membrane (PM) by preventing COPI binding to the cargo protein for anterograde transport.<sup>66,67</sup> Despite its low-complexity intracellular organization, *Giardia* possesses an efficient membrane trafficking system<sup>68</sup> and a complex microtubule (MT) cytoskeleton that is critically important for and deeply involved in intracellular protein trafficking.<sup>69</sup> In *Giardia* trophozoites the synthesis and export to PM of transmembrane-anchored variant surface antigens (VSPs) and cysteine-rich, non-variable protein, constitute a major part of the secretory activity. The transport of this cargo proteins is likely mediated by COPI, although a conventional post-ER/Golgi compartment has not been identified in *Giardia*



1  
2  
3 trophozoites.<sup>68</sup> In another protozoan parasite, *Trypanosoma brucei*, depletion of 14-3-3 results in the size  
4  
5 reduction of the Rab11-positive recycling endosome compartment, thus reducing the rate of return of  
6  
7 recycling Variant Surface Glycoprotein to the plasma membrane.<sup>70</sup> In *Giardia*, Rab11 co-localizes with ESVs  
8  
9 and actin microfilaments during encystation,<sup>71</sup> suggesting its involvement in ESV transport. Actin has been  
10  
11 also identified as a 14-3-3 interactor in this work, as well as in other studies.<sup>18,19</sup> Mammalian 14-3-3  
12  
13 isoforms have been implicated in the remodeling of the actin cytoskeleton in the priming phase of  
14  
15 exocytosis.<sup>72</sup> Hence, it may not be surprising that cytoskeleton components, principally tubulins, plus-end  
16  
17 and minus-end motor proteins (kinesin, dynein) and intraflagellar transport (IFT) machinery were co-  
18  
19 precipitated with g14-3-3. In addition to tubulins,<sup>13,17,19,39</sup> direct binding of 14-3-3 to kinesins has been  
20  
21 proved,<sup>11,12,15,18,47</sup> particularly for the mammalian KIF1C,<sup>47</sup> involved in retrograde vesicle transport from the  
22  
23 Golgi to the ER. In mammals, 14-3-3 binding to microtubule-associated proteins (MAPs) and several  
24  
25 microtubule-associated proteins kinases (MARKs)<sup>15</sup> has been proposed to affect the rate of microtubule  
26  
27 assembly. Knockdown experiments in the bloodstream-form of *T. brucei* have also demonstrated a pivotal  
28  
29 role of 14-3-3s in motility and cytokinesis.<sup>49</sup> Taken together, the identified proteins suggest that g14-3-3  
30  
31 could orchestrate a coordinated, bidirectional motility of cargo proteins and coated vesicles along the  
32  
33 cellular microtubules in *Giardia*.  
34  
35  
36  
37  
38  
39

40  
41 Based on our identification of key enzymes of the glycolytic/gluconeogenic pathway and of the  
42  
43 pyruvate metabolism, a role of g14-3-3 in energy metabolism is also emerging. *Giardia* lacks mitochondria  
44  
45 and relies on fermentative metabolism for energy.<sup>73</sup> Glucose, the main carbon source, is incompletely  
46  
47 catabolized to acetate, ethanol, alanine, and CO<sub>2</sub>, and the balance of end-product formation is sensitive to  
48  
49 O<sub>2</sub> tension and glucose concentration in the medium.<sup>73</sup> Furthermore, in *Giardia*, the synthesis of  
50  
51 phosphoenolpyruvate and pyruvate, the intermediate products of carbohydrate metabolism, occupy a  
52  
53 central node at the crossroads of several metabolic pathways. In plants 14-3-3 proteins down-regulate key  
54  
55 enzyme activities in carbohydrate metabolism, as suggested by the direct binding of 14-3-3 proteins to  
56  
57 trehalose-6-phosphate synthase, glyceraldehyde-3-phosphate dehydrogenase and ATP synthase<sup>9,74</sup> and by  
58  
59 knock-down experiments of six 14-3-3 proteins in potato plants, resulting in increased activity of sucrose  
60

1  
2  
3 phosphate synthase and starch synthase.<sup>75</sup> Similarly, in *S. cerevisiae* suboptimal 14-3-3 protein activity  
4  
5 leads to increased levels of proteins involved in carbon metabolism, especially those implicated in  
6  
7 gluconeogenesis.<sup>76</sup> In *Hydra vulgaris*, 14-3-3 proteins have been localized in food granules, suggesting that  
8  
9 they can perform metabolic functions, possibly by regulating enzymes involved in storing nutritional  
10  
11 compounds.<sup>17,77</sup> It is then reasonable to suppose that energy metabolism in *Giardia* needs to be regulated  
12  
13 in response to glucose, ATP and oxygen levels in a 14-3-3-mediated manner. This is also in accordance with  
14  
15 the identification herein of the giardial  $\alpha$ ,  $\beta$  and  $\gamma$  subunits of the V-ATPase. Cytosolic biosynthetic pathways  
16  
17 are heavy users of ATP and, in a number of organisms, mitochondrial and chloroplatic ATP synthases have  
18  
19 been reported to bind to and be regulated by 14-3-3,<sup>13,17,48</sup> which might act as a molecular switch from ATP  
20  
21 synthesis to ATP hydrolysis.<sup>48</sup>  
22  
23

24  
25  
26 Giardial 14-3-3 seems also involved in protein synthesis and degradation processes, on the basis of  
27  
28 the identification of giardial transcription/translation and ubiquitin/proteasome machinery components as  
29  
30 g14-3-3 interactors (Table 1). 14-3-3 has been previously reported to bind to ubiquitination and  
31  
32 deubiquitination enzymes. Among these, the ubiquitin ligases Nedd4-2,<sup>42</sup> which functions in regulating  
33  
34 epithelial sodium channel surface expression, and Cop1 (constitutive photomorphogenic 1), which  
35  
36 ubiquitylates p53 and affects DNA damage response,<sup>43</sup> and the mammalian deubiquitinating enzyme  
37  
38 UBPY/USP8.<sup>12,44</sup> Proteomic studies focusing on 14-3-3<sup>16,19</sup> as well as on *Trichoderma reesei* 26S  
39  
40 proteasome<sup>45</sup> showed that 14-3-3 proteins co-purify with the 19S regulatory cap and the 20S catalytic core  
41  
42 of the 26S proteasome in yeast, plants and mammals. Decreased levels of many proteins involved in amino  
43  
44 acid synthesis and translation have been shown in a yeast mutant strain with suboptimal 14-3-3 protein  
45  
46 activity.<sup>76</sup> However, a direct binding to a component of the translation machinery has been proved only for  
47  
48 the eukaryotic initiation factor 2 $\alpha$  (eIF2 $\alpha$ ) of the flatworm *Schistosoma mansoni*.<sup>41</sup> In our study, the giardial  
49  
50 UPF1, a conserved nonsense-mediated mRNA decay factor, co-purified with FLAG-g14-3-3. UPF1 has been  
51  
52 implicated in regulating the abundance not only of aberrant mRNA, but also of some naturally occurring  
53  
54 mRNAs in *Giardia*, including those for encystation-induced genes.<sup>78</sup> Nam7p, the yeast homolog of UFP1, has  
55  
56 been identified in a large scale proteomic analysis of *S. cerevisiae* 14-3-3 interactions *in vivo*,<sup>18</sup> while human  
57  
58  
59  
60



1  
2  
3 UPF1 interacts with SMG7, a protein with a 14-3-3-like domain.<sup>40</sup> The identification of a protein proven to  
4 affect encystation should be underscored, since g14-3-3 has been shown to exert a regulatory effect on  
5  
6 *Giardia* encystation.  
7

8  
9  
10 Based on our proteomic identifications, we directly prove that g14-3-3 forms a complex with gCDC7  
11 and a putative gDBF4 during both the trophozoite and encysting stages of the parasite. This is also the first  
12 evidence that a CDC7-DBF4 kinase complex (gDDK) occurs in *Giardia*. It is plausible that gDDK, as in other  
13 eukaryotes, might control DNA replication by co-regulating G1/S phase transition through phosphorylation  
14 and activation of the mini-chromosome maintenance protein complex (Mcm2-7 helicases). It would thus  
15 promote the assembly and firing of replication forks, is involved in chromatin structuring, and foster  
16 chromosome segregation during mitosis.<sup>56</sup> Studies of *Giardia* ploidy have highlighted the importance of  
17 DNA replication in the encystation process.<sup>59</sup> After encystation induction, the cell exits from cell cycle in  
18 the G2 stage, after a first round of DNA replication (two 4N nuclei) and during the end of encystation  
19 replicates its DNA again without an intervening cell division, a process termed endoreplication, giving rise  
20 to a cyst with four 4N nuclei.<sup>60</sup> The partial localization of gCDC7 in the nuclei during encystation is  
21 compatible with its function in DNA replication and/or chromosome segregation. In budding yeast, DDK  
22 controls multiple processes necessary to prepare the chromosomes for reductional segregation during  
23 meiosis I.<sup>79</sup> The inability to observe a similar nuclear staining in trophozoites may be due to the use of cells  
24 already in stationary phase (i.e. in the post-replicative G2 stage).<sup>59</sup> A number of proteins are retained in the  
25 cytosol as a consequence of 14-3-3 binding and their nuclear localization occurs following release from 14-  
26 3-3.<sup>36,63</sup> In the case of the interaction between g14-3-3 and gDDK we observe that these proteins form a  
27 stable complex during the trophozoite and encystation stages. However, we cannot exclude that g14-3-3  
28 affects the localization of gDDK. In fact, the shortening of g14-3-3 polyglycine chain, which we have  
29 previously shown to regulate the access of g14-3-3 to nuclei,<sup>23,24</sup> might allow the g14-3-3/gDDK complex to  
30 localize into the nuclei during encystation. On the other hand, the localization in the trophozoite of gCDC7  
31 to the flagellar pores might reflect a function of this region in the perception of external growth or  
32 differentiation stimuli, as reported for non-motile primary cilia.<sup>80</sup>  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 Besides the regulation of gDDK intracellular localization, g14-3-3 binding could mediate the  
4 interaction of gDDK with components of the pre-replication complex (pre-RC). In *S. cerevisiae* the 14-3-3  
5 isoform Bmh2 is required for normal entry into S phase and regulates the early steps of DNA replication by  
6 interacting with Mcm2 and Orc2, two members of the pre-RC. Furthermore, 14-3-3 together with the ORC  
7 (Origin Replication Complex) are essential for the binding of MCM proteins to replication origins during G1  
8 phase.<sup>38</sup> In several proteomic studies, 14-3-3 proteins were also found to interact directly with replication-  
9 initiation proteins, including mammalian Mcm3, Mcm5 and Mcm10.<sup>12, 13</sup> Noteworthy, we found here that  
10 the giardial MCM3 homolog co-purified with g14-3-3.  
11

12 A further indication of the role of g14-3-3 in regulation of *Giardia* cell cycle comes from the  
13 detection of interactions with the dual specificity protein phosphatase CDC14A, the G2/M-specific cyclin B  
14 and a CDK (cdc2/cdc28-like) protein kinase. The 14-3-3 protein has been demonstrated to act on the G2  
15 and M checkpoints. In fact, the mammalian isoform 14-3-3 $\sigma$  prevents cells from entering mitosis by  
16 sequestering the cyclin dependent kinase Cdc2/cyclin B1 complex in the cytoplasm.<sup>37</sup> Rad24, an *S. pombe*  
17 14-3-3 isoform, promotes mitotic exit and cytokinesis in late mitosis through binding and cytosolic  
18 retention of Clp1, a member of the Cdc14 family of phosphatases that dephosphorylates substrates of the  
19 cyclin-dependent kinase.<sup>36</sup> From a general point of view, in *Giardia* g14-3-3 might integrate energy sensing  
20 with growth and differentiation, similar to fission in yeast, where 14-3-3 proteins are deeply involved in  
21 switching from the mitotic cell cycle to sexual differentiation under nutrient starvation.<sup>31</sup> In fact, it has been  
22 proposed<sup>59</sup> that the differentiation of *Giardia* into cystic forms is reminiscent of meiosis, in which the  
23 genome is first replicated and then divided twice without DNA replication.  
24

25 The data presented herein lay the bases for further studies aimed to disclose the exact role of g14-  
26 3-3 in the processes highlighted in this work, both at mechanical and functional levels. Intriguingly, 15 of  
27 the identified proteins (Table 1) were also suggested as candidate targets for the development of novel  
28 anti-giardial drugs,<sup>22</sup> including the gCDC7 protein kinase. Thus, the large amount of data collected in this  
29 work is likely to lead to the identification of further putative therapeutic targets. Once direct g14-3-3-  
30 partner interactions are defined, it might become possible to design innovative drugs that target selected  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 g14-3-3 complexes, either promoting or disrupting their stability, thus enhancing the potency and  
4  
5 specificity of such compounds.  
6  
7 Finally, the present in vivo g14-3-3 interactome includes a large percentage of proteins whose  
8  
9 function is still unknown. Thus, this study represents an important step towards the elucidation of still  
10  
11 poorly characterized biological processes of this ancient parasite through functional proteomics  
12  
13 approaches.  
14  
15

#### 16 17 18 19 **Aknowledgments**

20  
21 We wish to thank Dr. Maria C. Touz, from the Instituto de Investigación Medica Mercedes y Martin  
22  
23 Ferreyra, CONICET, Cordoba, Argentina, for the kindly gift of the pTubApaH7-HApac vector, Dr. Adrian B.  
24  
25 Hehl, from the University of Zurich, Switzerland, for the anti-gCLH and anti-gβCOP antibodies, Dr. Staffan  
26  
27 Svard, from the University of Uppsala, Sweden, for the anti-CWP antibody, Dr. Henry van Keulen, from the  
28  
29 Cleveland State University, Ohio, USA, for the anti-gPGM antibody. This work was partially supported by the  
30  
31 European Commission (contract SANCO/2006/FOODSAFETY/032). SC and MC have carried out the mass  
32  
33 spectrometry experiments in the frame of the Telethon Proteomic Service (GTF08002).  
34  
35  
36  
37

#### 38 39 40 **Abbreviations**

41  
42 IPTG (isopropyl-thio-β-D-galactoside); RP-LC-MS/MS (nanoflow reversed-phase liquid chromatography  
43  
44 tandem mass spectrometry); FDR (false discovery rate); emPAI (Exponentially Modified Protein Abundance  
45  
46 Index); DAPI (4',6-diamidino-2-phenylindole); CLSM (Confocal Laser Scanning Microscopy); AOTF (Acousto-  
47  
48 Optical Tunable Filter); Blue Native PAGE (BN-PAGE); gCLH (chlatriin heavy chain); gβCOP (coatamer beta  
49  
50 subunit); gPGM (phosphoacetylglucosamine mutase); CWP (cyst wall protein); *gap1* (glyceraldehyde 3-  
51  
52 phosphate dehydrogenase 1); *gdh* (glutamate dehydrogenase); Dbf4 (Dumbbell-forming 4).  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

## Reference

- (1) Rosenquist, M.; Sehnke, P.; Ferl, R. J.; Sommarin, M.; Larsson, C. Evolution of the 14-3-3 protein family: does the large number of isoforms in multicellular organisms reflect functional specificity? *J. Mol. Evol.* **2000**, *51*, 446-458.
- (2) Aitken, A. 14-3-3 proteins: a historic overview. *Semin. Cancer Biol.* **2006**, *16*, 162-172.
- (3) Gardino, A.K.; Smerdon, S.J.; Yaffe, M.B. Structural determinants of 14-3-3 binding specificities and regulation of subcellular localization of 14-3-3-ligand complexes: a comparison of the X-ray crystal structures of all human 14-3-3 isoforms. *Semin. Cancer Biol.* **2006**, *16*, 173-182.
- (4) Muslin, A.J.; Tanner, J.W.; Allen, P.M.; Shaw, A.S. Interaction of 14-3-3 with signaling proteins is mediated by the recognition of phosphoserine. *Cell* **1996**, *84*, 889-897.
- (5) Yaffe, M.B.; Rittinger, K.; Volinia, S.; Caron, P.R.; Aitken, A.; Leffers, H.; Gamblin, S.J.; Smerdon, S. J.; Cantley, L.C. The structural basis for 14-3-3:phosphopeptide binding specificity. *Cell* **1997**, *91*, 961-971.
- (6) Coblitz, B.; Wu, M.; Shikano, S.; Li, M. C-terminal binding: an expanded repertoire and function of 14-3-3 proteins. *FEBS Lett.* **2006**, *580*, 1531-1535.
- (7) Petosa, C.; Masters, S.C.; Bankston, L.A.; Pohl, J.; Wang, B.; Fu, H.; Liddington, R.C. 14-3-3zeta binds a phosphorylated Raf peptide and an unphosphorylated peptide via its conserved amphipathic groove. *J. Biol. Chem.* **1998**, *273*, 16305-16310.
- (8) Hallberg, B. Exoenzyme S binds its cofactor 14-3-3 through a non-phosphorylated motif. *Biochem. Soc. Trans.* **2002**, *30*, 401-405.
- (9) Cotellet, V.; Meek, S.E.; Provan, F.; Milne, F.C.; Morrice, N.; MacKintosh, C. 14-3-3s regulate global cleavage of their diverse binding partners in sugar-starved Arabidopsis cells. *EMBO J.* **2000**, *19*, 2869-2876.
- (10) Milne, F.C.; Moorhead, G.; Pozuelo Rubio, M.; Wong, B.; Kulma, A.; Harthill, J.E.; Villadsen, D.; Cotellet, V.; MacKintosh, C. Affinity purification of diverse plant and human 14-3-3-binding partners. *Biochem. Soc. Trans.* **2002**, *30*, 379-381.

- 1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60
- (11) Jin, J.; Smith, F.D.; Stark, C.; Wells, C.D.; Fawcett, J.P.; Kulkarni, S.; Metalnikov, P.; O'Donnell, P.; Taylor, P.; Taylor, L.; Zougman, A.; Woodgett, J.R.; Langeberg, L.K.; Scott, J.D.; Pawson, T. Proteomic, functional, and domain-based analysis of in vivo 14-3-3 binding proteins involved in cytoskeletal regulation and cellular organization. *Curr. Biol.* **2004**, *14*, 1436-1450.
- (12) Meek, S.E.; Lane, W.S.; and Piwnica-Worms, H. Comprehensive proteomic analysis of interphase and mitotic 14-3-3-binding proteins. *J. Biol. Chem.* **2004**, *279*, 32046-32054.
- (13) Pozuelo-Rubio, M.; Geraghty, K.M.; Wong, B.H.; Wood, N.T.; Campbell, D.G.; Morrice, N.; Mackintosh, C. 14-3-3-affinity purification of over 200 human phosphoproteins reveals new links to regulation of cellular metabolism, proliferation and trafficking. *Biochem. J.* **2004**, *379*, 395-408.
- (14) Mackintosh, C. Dynamic interactions between 14-3-3 proteins and phosphoproteins regulate diverse cellular processes. *Biochem. J.* **2004**, *381*, 329-342.
- (15) Angrand, P.O.; Segura, I.; Völkel, P.; Ghidelli, S.; Terry, R.; Brajenovic, M.; Vintersten, K.; Klein, R.; Superti-Furga, G.; Drewes, G.; Kuster, B.; Bouwmeester, T.; Acker-Palmer, A. Transgenic mouse proteomics identifies new 14-3-3-associated proteins involved in cytoskeletal rearrangements and cell signaling. *Mol. Cell. Proteomics* **2006**, *12*, 2211-2227.
- (16) Alexander, R.D.; Morris, P.C. A proteomic analysis of 14-3-3 binding proteins from developing barley grains. *Proteomics* **2006**, *6*, 1886-1896.
- (17) Pauly, B.; Lasi, M.; MacKintosh, C.; Morrice, N.; Imhof, A.; Regula, J.; Rudd, S.; David, C.N.; Böttger, A. Proteomic screen in the simple metazoan Hydra identifies 14-3-3 binding proteins implicated in cellular metabolism, cytoskeletal organisation and Ca<sup>2+</sup> signalling. *BMC Cell. Biol.* **2007**, *8*, 31.
- (18) Kakiuchi, K.; Yamauchi, Y.; Taoka, M.; Iwago, M.; Fujita, T.; Ito, T.; Song, S. Y.; Sakai, A.; Isobe, T.; Ichimura, T. Proteomic analysis of in vivo 14-3-3 interactions in the yeast *Saccharomyces cerevisiae*. *Biochemistry* **2007**, *46*, 7781-7792.

- 1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60
- (19)Paul, A.L.; Liu, L.; McClung, S.; Laughner, B.; Chen, S.; Ferl, R.J. Comparative interactomics: analysis of arabidopsis 14-3-3 complexes reveals highly conserved 14-3-3 interactions between humans and plants. *J. Proteome Res.* **2009**, *8*, 1913-1924.
- (20)Johnson, C.; Crowther, S.; Stafford, M.J.; Campbell, D.G.; Toth, R.; MacKintosh, C. Bioinformatic and experimental survey of 14-3-3-binding sites. *Biochem. J.* **2010**, *427*, 69-78.
- (21)Thompson, R.C. Giardiasis as a re-emerging infectious disease and its zoonotic potential. *Int. J. Parasitol.* **2000**, *12-13*, 1259-1267.
- (22)Morrison, H.G.; McArthur, A.G.; Gillin, F.D.; Aley, S.B.; Adam, R.D.; Olsen, G.J.; Best, A.A.; Cande, W.Z.; Chen, F.; Cipriano, M.J.; Davids, B.J.; Dawson, S.C.; Elmendorf, H.G.; Hehl, A.B.; Holder, M.E.; Huse, S.M.; Kim, U.U.; Lasek-Nesselquist, E.; Manning, G.; Nigam, A.; Nixon, J.E.; Palm, D.; Passamanek, N.E.; Prabhu, A.; Reich, C.I.; Reiner, D.S.; Samuelson, J.; Svard, S.G.; Sogin, M.L. Genomic minimalism in the early diverging intestinal parasite *Giardia lamblia*. *Science* **2007**, *317*, 1921-1926.
- (23)Lalle, M.; Salzano, A.M.; Crescenzi, M.; and Pozio, E. The *Giardia duodenalis* 14-3-3 protein is post-translationally modified by phosphorylation and polyglycylation of the C-terminal tail. *J. Biol. Chem.* **2006**, *281*, 5137-5148.
- (24)Lalle, M.; Bavassano, C.; Fratini, F.; Cecchetti, S.; Boisguerin, P.; Crescenzi, M.; Pozio, E. Involvement of 14-3-3 protein post-translational modifications in *Giardia duodenalis* encystation. *Int. J. Parasitol.* **2010**, *40*, 201-213.
- (25)Lalle, M.; Camerini, S.; Cecchetti, S.; Fantauzzi, C. B.; Crescenzi, M.; Pozio, E. *Giardia duodenalis* 14-3-3 protein is polyglycylation by a tubulin tyrosine ligase-like member and deglycylation by two metalloproteinases. *J. Biol. Chem.* **2011**, *286*, 4471-4484.
- (26)Touz, M.C.; Lujan, H.D.; Hayes, S.F.; Nash, T.E. Sorting of encystation-specific cysteine protease to lysosome-like peripheral vacuoles in *Giardia lamblia* requires a conserved tyrosine-based motif. *J. Biol. Chem.* **2003**, *278*, 6420-6426.



- 1  
2  
3 (27)Winiiecka-Krusnell, J.; Linder, E. Detection of *Giardia lamblia* cysts in stool samples by  
4 immunofluorescence using monoclonal antibody. *Eur. J. Clin. Microbiol. Infect. Dis.* **1995**, *14*,  
5 218-222.  
6  
7  
8  
9  
10 (28)Marti, M. Regos, A.; Li, Y.; Schraner, E.M.; Wild, P.; Muller, N.; Knopf, L.G.; Hehl, A.B. An  
11 ancestral secretory apparatus in the protozoan parasite *Giardia intestinalis*, *J. Biol. Chem.* **2003**,  
12 278, 24837–24848.  
13  
14  
15 (29)Lopez, A.B.; Sener, K.; Jarroll, E.L.; van Keulen, H. Transcription regulation is demonstrated for  
16 five key enzymes in *Giardia intestinalis* cyst wall polysaccharide biosynthesis. *Mol. Biochem.*  
17 *Parasitol.* **2003**, *128*, 51-57.  
18  
19  
20  
21 (30)Shevchenko, A.; Wilm, M.; Vorm, O.; Mann, M. Mass spectrometric sequencing of proteins  
22 silver-stained polyacrylamide gels. *Anal. Chem.* **1996**, *68*, 850-858.  
23  
24  
25 (31)Yates, J.R. 3rd, Eng, J.K.; McCormack, A.L. Mining genomes: correlating tandem mass spectra of  
26 modified and unmodified peptides to sequences in nucleotide databases. *Anal. Chem.* **1995**, *67*,  
27 3202-3210.  
28  
29  
30  
31 (32)Ishihama, Y.; Oda, Y.; Tabata, T.; Sato, T.; Nagasu, T.; Rappsilver, J.; Mann, M. Exponentially  
32 modified protein abundance index (emPAI, for estimation of absolute protein amount in  
33 proteomics by the number of sequenced peptides per protein. *J. Mol. Cell. Proteomics.* **2005**, *4*,  
34 1265-1272.  
35  
36  
37 (33)Yee, J.; Tang, A.; Lau, W.L.; Ritter, H.; Delpont, D.; Page, M.; Adam, R.D.; Müller, M.; Wu, G. Core  
38 histone genes of *Giardia intestinalis*: genomic organization, promoter structure, and  
39 expression. *BMC Mol. Biol.* **2007**, *8*, 26.  
40  
41  
42 (34)Bustos, D.M.; Iglesias, A.A. Intrinsic disorder is a key characteristic in partners that bind 14-3-3  
43 proteins. *Proteins* **2006**, *63*, 35-42.  
44  
45  
46 (35)Galkin, A.; Kulakova, L.; Melamud, E.; Li, L.; Wu, C.; Mariano, P.; Dunaway-Mariano, D.; Nash, T.  
47 E.; Herzberg, O. Characterization, kinetics, and crystal structures of fructose-1,6-bisphosphate  
48 aldolase from the human parasite, *Giardia lamblia*. *J. Biol. Chem.* **2007**, *282*, 4859-4867.  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60



- 1  
2  
3  
4 (36)Chen, C.T.; Feoktistova, A.; Chen, J.S.; Shim, Y.S.; Clifford, D.M.; Gould, K.L.; McCollum, D. The  
5  
6 SIN kinase Sid2 regulates cytoplasmic retention of the *S. pombe* Cdc14-like phosphatase Clp1.  
7  
8 *Curr. Biol.* **2008**, *18*, 1594-1599.  
9  
10 (37)Laronga, C.; Yang, H.Y.; Neal, C.; Lee, M.H. Association of the cyclin-dependent kinases and 14-  
11  
12 3-3 sigma negatively regulates cell cycle progression. *J. Biol. Chem.* **2000**, *275*, 23106-23112.  
13  
14 (38)Yahyaoui, W.; Zannis-Hadjopoulos, M. 14-3-3 proteins function in the initiation and elongation  
15  
16 steps of DNA replication in *Saccharomyces cerevisiae*. *J. Cell. Sci.* **2009**, *122*, 4419-4426.  
17  
18 (39)Chang, I.F.; Curran, A.; Woolsey, R.; Quilici, D.; Cushman, J.C.; Mittler, R.; Harmon, A.; Harper,  
19  
20 J.F. Proteomic profiling of tandem affinity purified 14-3-3 protein complexes in *Arabidopsis*  
21  
22 *thaliana*. *Proteomics* **2009**, *9*, 2967-2985.  
23  
24 (40)Fukuhara, N.; Ebert, J.; Unterholzner, L.; Lindner, D.; Izaurrealde, E.; Conti, E. SMG7 is a 14-3-3-  
25  
26 like adaptor in the nonsense-mediated mRNA decay pathway. *Mol. Cell.* **2005**, *17*, 537-547.  
27  
28 (41)McGonigle, S.; Beall, M.J.; Pearce, E.J. Eukaryotic initiation factor 2 alpha subunit associates  
29  
30 with TGF beta receptors and 14-3-3 epsilon and acts as a modulator of the TGF beta response.  
31  
32 *Biochemistry* **2002**, *41*, 579-587.  
33  
34 (42)Nagaki, K.; Yamamura, H.; Shimada, S.; Saito, T.; Hisanaga, S.; Taoka, M.; Isobe, T.; Ichimura, T.  
35  
36 14-3-3 mediates phosphorylation-dependent inhibition of the interaction between the  
37  
38 ubiquitin E3 ligase Nedd4-2 and epithelial Na<sup>+</sup> channels. *Biochemistry* **2006**, *45*, 6733-6740.  
39  
40 (43)Su, C.H.; Zhao, R.; Velazquez-Torres, G.; Chen, J.; Gully, C.; Yeung, S.C.; Lee, M.H. Nuclear export  
41  
42 regulation of COP1 by 14-3-3 $\sigma$  in response to DNA damage. *Mol. Cancer* **2010**, *9*, 243.  
43  
44 (44)Mizuno, E.; Kitamura, N.; Komada, M. 14-3-3-dependent inhibition of the deubiquitinating  
45  
46 activity of UBPY and its cancellation in the M phase. *Exp. Cell. Res.* **2007**, *313*, 3624-3634.  
47  
48 (45)Grinyer, J.; Kautto, L.; Traini, M.; Willows, R.D.; Te'o, J.; Bergquist, P.; Nevalainen, H. Proteome  
49  
50 mapping of the *Trichoderma reesei* 20S proteasome. *Curr. Genet.* **2007**, *51*, 79-88.  
51  
52 (46)Pozuelo-Rubio, M. Regulation of autophagic activity by 14-3-3 $\zeta$  proteins associated with class III  
53  
54 phosphatidylinositol-3-kinase. *Autophagy* **2011**, *7*, 240-242.  
55  
56  
57  
58  
59  
60

- 1  
2  
3 (47)Dorner, C.; Ullrich, A.; Häring, H.U.; Lammers, R. The kinesin-like motor protein KIF1C occurs in  
4 intact cells as a dimer and associates with proteins of the 14-3-3 family. *J. Biol. Chem.* **1999**,  
5 274, 33654-33660.  
6  
7  
8  
9  
10 (48)Bunney, T.D.; van Walraven, H.S.; de Boer, A.H. 14-3-3 protein is a regulator of the  
11 mitochondrial and chloroplast ATP synthase. *Proc. Natl. Acad. Sci. U S A* **2001**, 98, 4249-4254.  
12  
13 (49)Inoue, M.; Yasuda, K.; Uemura, H.; Yasaka, N.; Inoue, H.; Sei, Y.; Horikoshi, N.; Fukuma, T.  
14 Phosphorylation-dependent protein interaction with *Trypanosoma brucei* 14-3-3 proteins that  
15 display atypical target recognition. *PLoS One* **2010**, 5, e15566.  
16  
17  
18 (50)Roberts, R.L.; Mösch, H.U.; Fink, G.R. 14-3-3 proteins are essential for RAS/MAPK cascade  
19 signaling during pseudohyphal development in *S. cerevisiae*. *Cell* **1997**, 89, 1055-1065.  
20  
21  
22 (51)Li, T.; Paudel, H.K. 14-3-3zeta facilitates GSK3beta-catalyzed tau phosphorylation in HEK-293  
23 cells by a mechanism that requires phosphorylation of GSK3beta on Ser9. *Neurosci. Lett.* **2007**,  
24 414, 203-208.  
25  
26  
27 (52)Lee, P.; Paik, S.M.; Shin, C.S.; Huh, W.K.; Hahn, J.S. Regulation of yeast Yak1 kinase by PKA and  
28 autophosphorylation-dependent 14-3-3 binding. *Mol. Microbiol.* **2011**, 79, 633-646.  
29  
30  
31 (53)Kligys, K.; Yao, J.; Yu, D.; Jones, J.C. 14-3-3zeta/tau heterodimers regulate Slingshot activity in  
32 migrating keratinocytes. *Biochem. Biophys. Res. Commun.* **2009**, 383, 450-454.  
33  
34  
35 (54)Ichimura, T.; Taoka, M.; Hozumi, Y.; Goto, K.; Tokumitsu, H. 14-3-3 Proteins directly regulate  
36 Ca<sup>2+</sup>/calmodulin-dependent protein kinase kinase alpha through phosphorylation-dependent  
37 multisite binding. *FEBS Lett.* **2008**, 582, 661-665.  
38  
39  
40 (55)Johnston, L.H.; Masai, H.; Sugino, A.A. Cdc7p-Dbf4p protein kinase activity is conserved from  
41 yeast to humans. *Prog Cell Cycle Res.* **2000**, 4, 61-69.  
42  
43  
44 (56)Labib, K. How do Cdc7 and cyclin-dependent kinases trigger the initiation of chromosome  
45 replication in eukaryotic cells? *Genes Dev.* **2010**, 24, 1208-1219.  
46  
47  
48 (57)Sawa, M.; Masai, H. Drug design with Cdc7 kinase: a potential novel cancer therapy target.  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

- 1  
2  
3 (58)Sato, N.; Arai, K.; Masai, H. Human and *Xenopus* cDNAs encoding budding yeast Cdc7-related  
4 kinases: in vitro phosphorylation of MCM subunits by a putative human homologue of Cdc7.  
5  
6  
7 *EMBO J.* **1997**, *16*, 4340-4351.  
8  
9  
10 (59)Bernander, R.; Palm, J.E.; Svärd, S.G. Genome ploidy in different stages of the *Giardia lamblia*  
11 life cycle. *Cell Microbiol.* **2001**, *3*, 55-62.  
12  
13  
14 (60)Reiner, D.S.; Ankarklev, J.; Troell, K.; Palm, D.; Bernander, R.; Gillin, F.D.; Andersson, J.O.; Svärd,  
15 S.G. Synchronisation of *Giardia lamblia*: identification of cell cycle stage-specific genes and a  
16 differentiation restriction point. *Int. J. Parasitol.* **2008**, *38*, 935-944.  
17  
18  
19 (61)Ogino, K.; Takeda, T.; Matsui, E.; Iiyama, H.; Taniyama, C.; Arai, K.; Masai, H. Bipartite binding of  
20 a kinase activator activates Cdc7-related kinase essential for S phase. *J. Biol. Chem.* **2001**, *276*,  
21 31376-31387.  
22  
23  
24 (62)Harkins, V.; Gabrielse, C.; Haste, L.; Weinreich, M. Budding yeast Dbf4 sequences required for  
25 Cdc7 kinase activation and identification of a functional relationship between the Dbf4 and  
26 Rev1 BRCT domains. *Genetics* **2009**, *183*, 1269-1282.  
27  
28  
29 (63)Nishino, T.G.; Miyazaki, M.; Hoshino, H.; Miwa, Y.; Horinouchi, S.; Yoshida, M. 14-3-3 regulates  
30 the nuclear import of class IIa histone deacetylases. *Biochem. Biophys. Res. Commun.* **2008**,  
31 377, 852-856.  
32  
33  
34 (64)Sonda, S.; Morf, L.; Bottova, I.; Baetschmann, H.; Rehrauer, H.; Caffisch, A.; Hakimi, M.A.; Hehl,  
35 A.B. Epigenetic mechanisms regulate stage differentiation in the minimized protozoan *Giardia*  
36 *lamblia*. *Mol. Microbiol.* **2010**, *76*, 48-67.  
37  
38  
39 (65)Rivero, M.R.; Vranych, C.V.; Bisbal, M.; Maletto, B.A.; Ropolo, A.S.; Touz, M.C. Adaptor protein 2  
40 regulates receptor-mediated endocytosis and cyst formation in *Giardia lamblia*. *Biochem. J.*  
41 **2010**, *428*, 33-45.  
42  
43  
44 (66)O'Kelly, I.; Butler, M.H.; Zilberberg, N.; Goldstein, S.A. Forward transport. 14-3-3 binding  
45 overcomes retention in endoplasmic reticulum by dibasic signals. *Cell* **2002**, *111*, 577-588.  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

- 1  
2  
3  
4 (67)Shikano, S.; Coblitz, B.; Sun, H.; Li, M. Genetic isolation of transport signals directing cell surface  
5 expression. *Nat. Cell Biol.* **2005**, *7*, 985-992.  
6  
7  
8 (68)Marti, M.; Li, Y.; Schraner, E. M.; Wild, P.; Köhler, P.; Hehl, A.B. The secretory apparatus of an  
9 ancient eukaryote: protein sorting to separate export pathways occurs before formation of  
10 transient Golgi-like compartments. *Mol. Biol. Cell* **2003a**, *14*, 1433-1447.  
11  
12  
13 (69)Dawson, S.C.; Sagolla, M.S.; Mancuso, J.J.; Woessner, D.J.; House, S.A.; Fritz-Laylin, L.; Cande,  
14 W.Z. Kinesin-13 regulates flagellar, interphase, and mitotic microtubule dynamics in *Giardia*  
15 *intestinalis*. *Eukaryot. Cell* **2007**, *6*, 2354-2364.  
16  
17  
18 (70)Benz, C.; Engstler, M.; Hillmer, S.; Clayton, C. Depletion of 14-3-3 proteins in bloodstream-form  
19 *Trypanosoma brucei* inhibits variant surface glycoprotein recycling. *Int. J. Parasitol.* **2010**, *40*,  
20 629-634.  
21  
22  
23 (71)Castillo-Romero, A.; Leon-Avila, G.; Wang, C.C.; Perez Rangel, A.; Camacho Nuez, M.; Garcia  
24 Tovar, C.; Ayala-Sumano, J.T.; Luna-Arias, J.P.; Hernandez, J.M. Rab11 and actin cytoskeleton  
25 participate in *Giardia lamblia* encystation, guiding the specific vesicles to the cyst wall. *PLoS*  
26 *Negl. Trop. Dis.* **2010**, *4*, e697.  
27  
28  
29 (72)Chamberlain, L.H.; Roth, D.; Morgan, A.; Burgoyne, R.D. Distinct effects of alpha-SNAP, 14-3-3  
30 proteins, and calmodulin on priming and triggering of regulated exocytosis. *J. Cell. Biol.* **1995**,  
31 130, 1063-1070.  
32  
33  
34 (73)Adam, R.D. Biology of *Giardia lamblia*. *Clin. Microbiol. Rev.* **2001**, *14*, 447-475.  
35  
36  
37 (74)Moorhead, G.; Douglas, P.; Cotelle, V.; Harthill, J.; Morrice, N.; Meek, S.; Deiting, U.; Stitt, M.;  
38 Scarabel, M.; Aitken, A.; MacKintosh, C. Phosphorylation-dependent interactions between  
39 enzymes of plant metabolism and 14-3-3 proteins. *Plant J.* **1999**, *18*, 1-12.  
40  
41  
42 (75)Zuk, M.; Weber, R.; Szopa, J. 14-3-3 protein down-regulates key enzyme activities of nitrate and  
43 carbohydrate metabolism in potato plants. *J. Agric. Food Chem.* **2005**, *53*, 3454-3460.  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

- (76) Bruckmann, A.; Hensbergen, P.J.; Balog, C.I.; Deelder, A.M.; de Steensma, H.Y.; van Heusden, G.P. Post-transcriptional control of the *Saccharomyces cerevisiae* proteome by 14-3-3 proteins. *J. Proteome Res.* **2007**, *5*, 1689-1699.
- (77) Pauly, B.; Stiening, B.; Schade, M.; Alexandrova, O.; Zoubek, R.; David, C.N.; Böttger, A. Molecular cloning and cellular distribution of two 14-3-3 isoforms from Hydra: 14-3-3 proteins respond to starvation and bind to phosphorylated targets. *Exp. Cell Res.* **2003**, *285*, 15-26.
- (78) Chen, Y.H.; Su, L.H.; Huang, Y.C.; Wang, Y.T.; Kao, Y.Y.; Sun, C.H. UPF1, a conserved nonsense-mediated mRNA decay factor, regulates cyst wall protein transcripts in *Giardia lamblia*. *PLoS One* **2008**, *10*, e3609.
- (79) Matos, J.; Lipp, J.J.; Bogdanova, A.; Guillot, S.; Okaz, E.; Junqueira, M.; Shevchenko, A.; Zachariae, W. Dbf4-dependent CDC7 kinase links DNA replication to the segregation of homologous chromosomes in meiosis I. *Cell* **2008**, *135*, 662-678.
- (80) Eggenschwiler, J.T.; Anderson, K.V. Cilia and developmental signaling. *Annu. Rev. Cell Dev. Biol.* **2007**, *23*, 345-373.
- (81) Oowatari, Y.; Toma, K.; Ozoe, F.; Kawamukai, M. Identification of sam4 as a rad24 allele in *Schizosaccharomyces pombe*. *Biosci. Biotechnol. Biochem.* **2009**, *73*, 1591-1598.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

**Figure captions:**

Fig.1 Expression in *Giardia duodenalis* of the FLAG-g14-3-3 and affinity isolation of the g14-3-3-associated proteins. A) g14-3-3 promoter region. AT-rich sequence is grey boxed and three putative g-CAB elements are black boxed. Nucleotide numbering refers to the first nucleotide of the starting codon (ATG in bold). B) Immunodetection of FLAG-Pb14-3-3. Twenty  $\mu\text{g}$  of soluble protein extracts from trophozoite (T) or parasite after 12h in encysting medium (12h) of the *G. duodenalis* parental WB-C6 strain or the transgenic FLAG-g14-3-3 line, were blotted and probed with mouse  $\alpha$ -FLAG or  $\alpha$ -g14-3-3. Black bracket indicates the bands corresponding to the FLAG-g14-3-3, whereas the grey bracket indicates the endogenous g14-3-3 with long (upper band) or short (lower band) polyglycine chain. Encystation was checked by immunoblotting with anti-phosphoacetylglucosamine mutase Ab ( $\alpha$ -gPGM). C) Immunoprecipitation of the FLAG-g14-3-3 protein complexes using  $\alpha$ -FLAG-conjugated beads. FLAG peptide-eluted complexes co-immunoprecipitated with the FLAG-g14-3-3 from *G. duodenalis* WB-C6 or FLAG-g14-3-3 expressing trophozoite (T) or parasite after 12h in encysting medium (12h) were run on a 4-12% SDS-PAGE. Gel was either stained with coomassie (first panel) or immunoblotted with the mouse  $\alpha$ -g14-3-3 (second panel) or subjected to overlay assay with the recombinant GST-g14-3-3 (third panel), either in absence or presence of 100  $\mu\text{M}$  of the A8Ap phosphopeptide (fourth panel). Asterisks indicate the protein bands corresponding to the FLAG-g14-3-3 and endogenous g14-3-3.

Fig.2 Confirmation of interactions detected by proteomic analysis. *Giardia* proteins co-immunoprecipitated with the FLAG-g14-3-3 and eluted by the FLAG peptide from trophozoite (T) or parasite after 12h in encysting medium (12h) from the *G. duodenalis* WB-C6 or FLAG-g14-3-3 lines were run on a 4-12% SDS-PAGE and immunoblotted with the indicated antibodies: gCLH (chlatrin heavy chain), anti-g $\beta$ COP (coatamer beta) and anti-gPGM (phosphoacetylglucosamine mutase).



1  
2  
3 Fig.3 Expression and localization of the gCDC7 during trophozoite and encystation stages of *G. duodenalis*  
4 WB-C6 strain. A) Expression of *gcdc7* mRNA. Quantitative real-time PCR analysis of *gcdc7* and *cwp1* (cyst  
5 wall protein 1) genes amplified with gene specific primers at different time points during the encystation (0,  
6 3, 6 and 12h). The threshold cycle of all genes was normalized to that of glyceraldehyde-3-phosphate  
7 dehydrogenase (*gap1*). Relative gene expression was calculated using the  $2^{-\Delta\Delta Ct}$  method. B) Stage-  
8 dependent expression of gCDC7 protein. Twenty  $\mu$ g of soluble proteins both from trophozoites and  
9 encysting parasites (induced at the indicated times) were used in each lane and immunoblotted with anti-  
10 gCDC7 polyclonal serum. The equal loading of the samples is checked by immunostaining with the anti- $\alpha$ -  
11 tubulin (panel  $\alpha$ TUB). C) Subcellular localization of the gCDC7 protein in *G. duodenalis* parasites. CLSM  
12 observations of trophozoites (Troph.), encysting parasites (at 12h post encystations induction, 12h Encyst.)  
13 and cysts stained with anti-gCDC7 polyclonal serum revealed by Alexa Fluor-594 goat anti-mouse Ab  
14 (shown in red), with anti-g14-3-3 serum or anti- $\alpha$ -tubulin ( $\alpha$ TUB), revealed by Alexa Fluor-488 goat anti-  
15 rabbit Ab (green), with Cy3-conjugated anti-CWP mAb (pseudocolor grey) and with DAPI (blue). Displayed  
16 micrographs correspond to single stack. Merged+DAPI, merged images with DAPI-stained nuclei. 3D, three  
17 dimensional reconstructions of the complete stack series for each acquisition. T, transmission light  
18 acquisition. Scale bars 2  $\mu$ m. In the trophozoite triangles indicate the position of exemplar flagellar pores,  
19 arrows indicate nuclei (N) and median body (mb). D). Western blot analysis of subcellular fractions of *G.*  
20 *duodenalis* trophozoites and parasite after 12h of growth in encystation medium. Twenty  $\mu$ g of the soluble  
21 (S) and membrane (M) fractions were separated by 4-12% SDS-PAGE, transferred to a PVDF membrane and  
22 immunostained with the indicated antibodies. Molecular size markers (kDa) are on the left.  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52 Fig.4 FLAG-gCDC7 protein expressed in *G. duodenalis* form a complex with g14-3-3 and gDBF4. A)  
53 Expression profile of the FLAG-tagged recombinant protein. Twenty  $\mu$ g of protein lysate from trophozoite  
54 (T) or from parasite after 12h in encysting medium (12h) of the *G. duodenalis* control strain WB-C6 or the  
55 FLAG-gCDC7 transfected parasite were separated on 4-12% SDS-PAGE. Immunoblots were performed with  
56 the indicated antibodies. Molecular size markers (kDa) are on the left. The equal loading of the sample  
57  
58  
59  
60



1  
2  
3 material is demonstrated by immunostaining with the anti- $\alpha$ -tubulin (panel  $\alpha$ TUB). B) Immunopurification  
4 of the FLAG-gCDC7 protein. An aliquot (1/20) of the FLAG peptide-eluted material (WB-C6 or FLAG-gCDC7  
5 transfectant) from trophozoite (T) or from parasite after 12h in encysting medium (12h) was separated by  
6 4-12% SDS-PAGE and immunoblotted with anti-gCDC7 or anti-g14-3-3 polyclonal sera. Position of the g14-  
7 3-3 protein with long polyglycine chain (upper band) or short polyglycine chain (lower band) is indicated by  
8 triangles. C) Overlay assay. An aliquot (1/20) of FLAG-immunoprecipitated (IP) from FLAG-gCDC7  
9 transfectant trophozoite, or control immunopurification from WB-C6 strain, were separated on 12% SDS-  
10 PAGE, coomassie stained (left panel) or transferred on nitrocellulose and incubated with the recombinant  
11 GST-g14-3-3 (right panel). The asterisk indicates the two band recognized by GST-g14-3-3 only in the FLAG-  
12 gCDC7 IP. D) Identification of the g14-3-3/gCDC7/gDBF4 complex. FLAG peptide-eluted immunoprecipitated  
13 material from trophozoite (T) or from parasite after 12h in encysting medium (12h) of WB-C6 strain, FLAG-  
14 g14-3-3 and FLAG-gCDC7 transfectants were separated on 3-12% Blue Native-PAGE and silver stained.  
15 Arrow indicates the band containing the g14-3-3/gCDC7/gDBF4 complex as revealed by nanoflow reversed-  
16 phase liquid chromatography tandem mass spectrometry. Molecular size markers (kDa) are indicated on  
17 the left of each panel.  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

**Tables 1.** Putative FLAG-g14-3-3 interacting proteins identified by combining FLAG-IP and RP-LC-MS/MS.

Protein identified by FLAG-IP and RP-LC-MS/MS were grouped into functional classes. FLAG-g14-3-3 associated proteins from FLAG-g14-3-3 transfected trophozoite (T) and from transfected parasite after 12h in encystation medium (12h). Accession number correspond to the gene ID at the Giardia DB (<http://giardiadb.org/>). Gray boxes indicate proteins identified in both trophozoite and encysting parasite samples; while black boxes indicate proteins identified only in encysting parasite samples.

	Product Description <sup>a, b, c</sup>	Accession number	MW	Peptide identified <sup>d</sup>		References <sup>e</sup>
				T	12h	
1						
2	<b>Cell cycle</b>					
3	Dual specificity protein phosphatase CDC14A (CDC14-like) <sup>f</sup>	GL50803_9270	66282	X		36
4	G2/M-specific cyclin B	GL50803_3977	39544	X		
5	Kinase, CMGC CDK (Cdc2-like CDK2/CDC28-like protein kinase) <sup>g</sup>	GL50803_16802	32863	X		37
6	Kinase, CDC 7 <sup>h</sup>	GL50803_112076	189973	X	X	
7	Hypothetical protein with zf-DBF domain to associate with CDC7 (Dbf4-homolog)	GL50803_94117	241085	X	X	
8	MCM3	GL50803_16214	102100	X		11-13, 38
9	<b>DNA, Chromatin and Chromosome Processing</b>					
10	Endonuclease III	GL50803_3595	36186	X		
11	ATP-binding protein (XBA1 GTPase-like protein)	GL50803_15164	42130	X		
12	Histone H2B	GL50803_121045	14590	X		12-13, 19
13	Golgi/cell cycle associated protein NUF2 (NDC80 kinetochore complex component)	GL50803_17472	54987	X		
14	Chromodomain helicase-DNA-binding protein, putative	GL50803_112978	300504	X		
15	DNA helicase TIP49 (RuvB-like protein 1/tpotin 52)	GL50803_9825	51419	X		
16	Hypothetical protein, similar to SMC2 (Structural Maintenance of Chromosomes)	GL50803_23185	173487	X		
17	Hypothetical protein (with SMC domain)	GL50803_14963	96761	X		
18	Hypothetical protein (with SMC domain)	GL50803_112112	136315	X		
19	Hypothetical protein (with SMC domain)	GL50803_17403	174283	X		
20	Hypothetical protein (with SMC domain)	GL50803_32999	51468	X		
21	Hypothetical protein (chromosome segregation protein; Provisional)	GL50803_6886	102837	X		
22	Hypothetical protein (with SbcC-ATPase involved in DNA repair/SMC domain)	GL50803_5883	62565	X		
23	Coiled-coil protein (with SMC domain)	GL50803_16332	196413	X		
24	Coiled-coil protein (with SbcC/SMC domain)	GL50803_14345	160223	X		
25	Coiled-coil protein (with SMC domain, similar to centriolin)	GL50803_42196	112970	X		
26	Coiled-coil protein (with SbcC/SMC domain)	GL50803_16342	186418	X		
27	Coiled-coil protein (with chromosome segregation ATPase domain)	GL50803_9492	107421	X		
28	Hypothetical protein (with SMC domains)	GL50803_113480	138931	X		
29	Hypothetical protein (with SMC domain)	GL50803_16630	140636	X		
30	Hypothetical protein (chromosome segregation protein; Provisional)	GL50803_95406	77851	X	X	
31	Hypothetical protein (with SMC domain)	GL50803_16068	84764	X	X	
32	<b>Transcription</b>					
33	DNA-directed RNA polymerase RPB3 RNA polymerase II subunit C	GL50803_7474	36100	X		39
34	DNA-directed RNA polymerase RPB1 RNA polymerase II subunit A	GL50803_89347	230457	X		
35	Transcription factor TFIIS	GL50803_5158	39179	X		
36	DRE4 protein (FACT complex subunit SPT16-like) <sup>h</sup>	GL50803_17430	129762		X	
37	Hypothetical protein (with TATA binding and SMC domain)	GL50803_92760	54718	X		
38	<b>RNA metabolism</b>					
39	ATP-dependent RNA helicase p54, putative (DDX6/DHH1)	GL50803_2098	48438	X		
40	ATP-dependent RNA helicase-like protein (DEAD box RNA helicase)	GL50803_15048	73380	X		18-19
41	DEAD box RNA helicase Vasa	GL50803_34684	50101	X		
42	ATP-dependent RNA helicase (DEAD box RNA helicase)	GL50803_13220	84457	X		
43	Hypothetical protein, (DEAD box RNA helicase required for poly(A <sup>+</sup> ) mRNA export)	GL50803_6283	53928	X		
44	RNA binding protein, putative	GL50803_11186	31501	X		
45	Protein BAP28	GL50803_103285	262513	X		
46	PNO1 (Partner of Nob1) <sup>i</sup>	GL50803_16718	22363	X		18
47	Nucleolar GTPase	GL50803_16498	60211	X		
48	Regulator of nonsense transcripts 1-like protein (UPF1)	GL50803_13452	147484		X	18, 40
49	Hypothetical protein (with RNA recognition motif)	GL50803_7204	18594	X		
50	Hypothetical protein (with K homology RNA-binding domain)	GL50803_8485	95821	X		
51	Hypothetical protein (with CCR4-Not complex component domain)	GL50803_96732	383737	X		13
52	<b>Translation</b>					
53	Ribosomal protein S2	GL50803_8118	26724	X		11
54	Ribosomal protein S3	GL50803_7999	24738	X		12, 18
55	Ribosomal protein S5	GL50803_12981	21025	X		
56	Ribosomal protein S6	GL50803_14620	28028	X		
57	Ribosomal protein S9	GL50803_4547	21654	X	X	
58	Ribosomal protein S12 <sup>j</sup>	GL50803_33862	14354	X		
59	Ribosomal protein S13	GL50803_16652	17591	X		
60	Ribosomal protein S14 <sup>k</sup>	GL50803_7678	15726	X		
61	Ribosomal protein S15A	GL50803_15228	14749	X		
62	Ribosomal protein S17 <sup>l</sup>	GL50803_6135	15726	X		
63	Ribosomal protein S23	GL50803_14699	15864	X		
64	Ribosomal protein S24	GL50803_10367	14844	X		
65	Ribosomal protein L2	GL50803_16086	27038	X		39
66	Ribosomal protein L9	GL50803_17056	20856	X	X	
67	Ribosomal protein L12	GL50803_14938	19572	X		
68	Ribosomal protein L13A	GL50803_11247	22873	X	X	
69	Ribosomal protein L14	GL50803_14091	14781	X		
70	Ribosomal protein L15	GL50803_8001	24295	X		
71	Ribosomal protein L23	GL50803_10081	15422	X		
72	Ribosomal protein L24A	GL50803_19003	11976	X		
73	Ribosomal protein L30 <sup>m</sup>	GL50803_14321	11685	X		
74	Ribosomal protein L35A <sup>n</sup>	GL50803_5947	13889	X	X	
75	Asparaginyl-tRNA synthetase	GL50803_14375	62439	X		
76	Cysteinyl-tRNA synthetase <sup>o</sup>	GL50803_5867	72994	X		
77	Tyrosyl-tRNA synthetase <sup>p</sup>	GL50803_16612	43487	X		
78	Hypothetical protein (Tyrosyl-tRNA synthetase)	GL50803_17393	37916	X		
79	Arginyl-tRNA synthetase <sup>q</sup>	GL50803_10521	70296	X		

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

Glutamyl-tRNA synthetase <sup>a</sup>	GL50803_9348	87603	X			
Glutamyl-tRNA synthetase	GL50803_86681	80131	X			12, 39
Prolyl-tRNA synthetase	GL50803_15983	61019	X	X		
Alanyl-tRNA synthetase	GL50803_96460	108440	X			12
Hypothetical protein, queuine tRNA-ribosyltransferase <sup>b</sup>	GL50803_6531	42233	X			
Translation initiation factor eIF2 gamma subunit	GL50803_2970	52324	X			12, 19, 41
Translation initiation factor eIF-2B alpha subunit	GL50803_91911	36736	X			19, 39
Hypothetical protein, Translation initiation factor 3 (eIF-3 subunit 9)	GL50803_15495	98490	X			12
Elongation initiation factor 5C	GL50803_7522	46894	X			17, 19
<b>Protein degradation</b>						
Ubiquitin-protein ligase E3A	GL50803_17386	131415	X			
Hypothetical protein (Ubiquitin-protein ligase E3A-like)	GL50803_137754	107853	X			12, 18, 42-43
Hypothetical protein (with HECT domain)	GL50803_32730	187850	X			
Ubiquitin carboxyl-terminal hydrolase 4	GL50803_14460	93056	X			
Ubiquitin carboxyl-terminal hydrolase 14	GL50803_8189	50497	X			12, 44
Ubiquitin carboxyl-terminal hydrolase 14	GL50803_102710	91751		X		
26S proteasome, 19S non-ATPase regulatory subunit 2 (RPN1)	GL50803_33166	135892	X			
26S proteasome, 19S regulatory subunit (RPN2)	GL50803_91643	147833	X			16, 19, 45
26S proteasome, 19S regulatory subunit 8 (RPT8)	GL50803_17106	45037	X			
26S proteasome, 19S regulatory subunit 6A (RPT5)	GL50803_4365	55970	X			
26S proteasome, 19S regulatory subunit 6B (RPT3)	GL50803_7950	43352	X			
Hypothetical protein, 26S proteasome, 19S regulatory subunit (RPN10)	GL50803_15604	28355	X			
Methionine aminopeptidase <sup>c</sup>	GL50803_86600	47108	X			
Aminoacyl-histidine dipeptidase (gDIP2)	GL50803_8407	55104	X			
Metalloprotease, insulinase family, family M16C (clan ME)	GL50803_9508	131440	X			
<b>Transport</b>						
Arsenical pump-driving ATPase (Get3-homolog)	GL50803_7953	39506	X			
Hypothetical protein, Transmembrane MFS general substrate transporter putative	GL50803_112063	72525	X			
Phosphatidylinositol transfer protein alpha isoform (membrane-associated)	GL50803_4197	39316	X			
<b>Ribosome to ER</b>						
Signal recognition particle receptor alpha subunit (SRPRalpha)	GL50803_14856	62712	X			19
SecE1-alpha	GL50803_5744	54093	X			16
<b>Nucleo/Cytoplasm</b>						
Hypothetical protein, with Cse1 domain (nuclear export of importin alpha)	GL50803_17110	108810	X			12
Hypothetical protein, exportin-1 (CRM1-mediated nuclear export)	GL50803_93278	119014	X			
Serologically defined colon cancer antigen 1 (nuclear export mediator)	GL50803_4043	118900		X		
<b>Membrane fusion</b>						
NSF (N-ethylmaleimide-sensitive fusion protein, vesicle-fusing ATPase), Sec18-like	GL50803_112681	90659	X	X		39
NSF, Sec18-like	GL50803_114776	89605	X			
ARF3 (ADP ribosylation factor/Ras-like GTPase family) <sup>d</sup>	GL50803_13930	20397	X			
Hypothetical protein (with Sec7 domain, GEF)	GL50803_112258	248743	X			11, 16
Rab2a (GTP-binding/Ras small GTPase) exocytic	GL50803_15567	23749	X			
Rab11 (GTP-binding/Ras small GTPase) recycling	GL50803_1695	23571	X			15, 19
Rab GDI	GL50803_11495	52597	X	X		
<b>Vesicle transport</b>						
COPI vesicle coatamer complex-alpha subunit	GL50803_11953	139329	X			
COPI vesicle coatamer complex-beta subunit	GL50803_88082	114880	X			
COPI vesicle coatamer complex-gamma subunit	GL50803_5603	105262	X			
COPI vesicle coatamer complex-delta subunit	GL50803_6170	32891	X			
Sec23 (COPII complex)	GL50803_9376	95209	X			13, 16
Claflrin heavy chain	GL50803_102108	206956	X	X		11-13
Alpha adaptin (AP2 subunit)	GL50803_17304	87036	X			
Beta adaptin (AP2 subunit)	GL50803_21423	122833	X	X		
Mu adaptin (AP2 subunit)	GL50803_8917	48415	X			18
Transitional endoplasmic reticulum ATPase (VPS4-like)	GL50803_8524	47143	X			
VPS15, Myristoylated serine/threonine protein kinase <sup>e</sup>	GL50803_113456	234044	X			
VPS34-like Phosphoinositide-3-kinase, class 3 <sup>f</sup>	GL50803_17406	183282	X			46
VPS35, Vacuolar protein sorting 35	GL50803_23833	87506	X	X		
Hypothetical protein, ESCRT-II complex subunit VPS22	GL50803_90710	27299	X			
Hypothetical protein, ESCRT-II complex subunit VPS25	GL50803_8329	21416	X			
Protein 21.1 with ANK and SMC domains	GL50803_103810	206466	X			
<b>Cytoskeletal components and associated proteins</b>						
Alpha-14 giardin <sup>g</sup>	GL50803_15097	38585	X			
Gamma giardin	GL50803_17230	35633	X			
Tubulin alpha I	GL50803_103676	50553	X	X		13, 17, 19, 39
Tubulin beta I	GL50803_101291	50050	X	X		
Tubulin Tyrosin Ligase-Like 3 (gTLL3)	GL50803_8456	74068	X			
Tubulin specific chaperone D	GL50803_10145	144480	X			
TCP-1 chaperonin subunit alpha	GL50803_91919	59282	X			
TCP-1 chaperonin subunit alpha	GL50803_17482	56325	X			
TCP-1 chaperonin subunit beta	GL50803_11397	56605	X			
TCP-1 chaperonin subunit gamma	GL50803_17411	61561	X	X		11, 19;
TCP-1 chaperonin subunit epsilon	GL50803_11992	61194	X	X		
TCP-1 chaperonin subunit eta	GL50803_16124	64754	X			
TCP-1 chaperonin subunit theta	GL50803_13500	60647	X			
TCP-1 chaperonin subunit zeta	GL50803_10231	60942	X			
Actin	GL50803_40817	41587	X			18-19
Hypothetical protein with Ca2+-binding protein EF-Hand superfamily (actin related protein)	GL50803_8727	124484	X			
Hypothetical protein (with myosin heavy chain domain)	GL50803_13651	78445	X			
STU2-like protein, (Microtubule-associated protein of the XMAP215/Dis1 family)	GL50803_91480	187774	X			
Hypothetical protein with Rhodanese Homology Domain (similar to Cep41)	GL50803_5012	42014	X			

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

WD-repeat membrane protein (Axoneme associated FAP66)	GL50803_16709	213474	X		
Tubby superfamily protein (Axoneme associated FAP118)	GL50803_87817	169399	X		
Axoneme central apparatus protein	GL50803_16202	54245	X		
Dynein heavy chain (axonemal DNAH5/DNAH8 homolog)	GL50803_17265	302122	X		
Dynein heavy chain (axonemal DNAH5/DNAH8 homolog)	GL50803_103059	274111	X		
Dynein heavy chain (axonemal DNAH6 homolog)	GL50803_111950	570325	X		12
Dynein heavy chain (axonemal DNAH9 homolog)	GL50803_16994	292007	X		
Dynein heavy chain (axonemal DNAH9 homolog)	GL50803_8172	43330	X		
Dynein heavy chain, putative (axonemal DNAH10 homolog)	GL50803_100906	623003	X		
Dynein intermediate chain (Flagellar outer dynein arm intermediate chain, DIC78)	GL50803_33218	83864	X		
Dynein heavy chain 11 ciliary	GL50803_42285	834752	X		
Dynein light chain (cytoplasmic DNLC8)	GL50803_9848	10448	X		
Median body protein	GL50803_16343	100585	X		
Kinesin-1	GL50803_13825	107938	X		
Kinesin-like protein	GL50803_14070	87005	X		
Hypothetical protein (with Kinesin motor domain and SMC domain)	GL50803_16518	121737	X		11, 12, 15, 18, 47
Spindle pole protein, putative (kinesin family member 5C)	GL50803_18930	105302	X	X	
Hypothetical protein (trafficking kinesin binding 1 homology)	GL50803_32676	80926	X		
Tubulin gamma ring complex component	GL50803_12057	101158	X		39
SALP-1 (striated fiber-assemblin family)	GL50803_4410	29828	X		
IFT cytoplasmic Dynein heavy chain 1b (cytoDHC-1b)	GL50803_93736	523076	X		
IFT complex A (IFT140-like)	GL50803_17251	212381	X		
IFT complex A (IFT122-like)	GL50803_16547	188006	X		
IFT complex B (IFT88)	GL50803_16660	95987	X		
IFT complex B associated DYF-3 (Clusterin associated protein 1, putative)	GL50803_16707	48630	X		
TRP domain containing protein	GL50803_11100	200789	X	X	
<b>Energy and Metabolism</b>					
<b>Fatty acid and lipids metabolism</b>					
Acetyl-CoA carboxylase (ACC)/pyruvate carboxylase fusion protein, putative	GL50803_113021	148197	X		13, 18, 19, 39
Hypothetical protein (with Long chain fatty acid CoA ligase-like domain)	GL50803_17174	212737	X		12
Hypothetical protein, glycerophosphodiester phosphodiesterase (GDPD)	GL50803_6492	68555	X		
Glycerol-3-phosphate dehydrogenase	GL50803_16125	118855	X	X	
Inositol-3-phosphate synthase	GL50803_17579	59906	X	X	39
Farnesyl diphosphate synthase	GL50803_6633	46232	X		
<b>Aminoacids metabolism</b>					
L-serine dehydratase SD1	GL50803_24662	57068	X		
Alanine aminotransferase, putative	GL50803_16363	53151	X		
5-methylthioadenosine nucleosidase, S-adenosylhomocysteine nucleosidase	GL50803_20195	28908	X		
<b>Purine and Pyrimidine metabolism</b>					
CTP synthase/UTP-ammonia lyase	GL50803_4507	99352	X		11-12
Uridine kinase	GL50803_8217	66045	X		
Nicotinate phosphoribosyltransferase	GL50803_9038	69673	X		
<b>Carbohydrates metabolism</b>					
Ribulose-phosphate 3-epimerase (Pentose-5-phosphate 3-epimerase or PPE)	GL50803_10324	25703	X		
Malic enzyme [NADP malate dehydrogenase (decarboxylating)]	GL50803_14285	61600	X	X	19
Phosphoacetylglucosamine mutase	GL50803_16069	56556	X	X	
<b>Glycolysis/Gluconogenesis</b>					
Glucokinase, lateral transfer candidate	GL50803_8826	37713	X		
UTP-glucose-1-phosphate uridylyltransferase	GL50803_29307	49294	X	X	
Glycogen synthase, putative	GL50803_104031	85063	X	X	
Fructose-bisphosphate aldolase	GL50803_11043	35214	X		17, 19
Glyceraldehyde 3-phosphate dehydrogenase (gap1) <sup>3</sup>	GL50803_6687	36336	X	X	9, 13, 16-18, 39
Pyruvate kinase	GL50803_3208	70739	X		
Pyruvate kinase	GL50803_17143	60526	X		19
<b>Formate/Ethanol/Acetate metabolism</b>					
Iron-containing Alcohol dehydrogenase 3, lateral transfer candidate <sup>3</sup>	GL50803_3861	45227	X		16, 19
Acetyl-CoA synthetase (ADP-forming)	GL50803_13608	78101	X	X	39
Pyruvate-formate lyase-activating enzyme, lateral transfer candidate	GL50803_9368	38919	X		
<b>Energy</b>					
V-ATP synthase subunit A, putative	GL50803_18470	104148	X		
V-ATP synthase subunit B	GL50803_12216	54747	X	X	13, 17, 48
V-ATP synthase subunit E, putative	GL50803_13603	23381	X		
<b>Protein kinase and phosphatase</b>					
PP2A Ser/Thr phosphatase, 65kDa PP2A regulatory subunit A	GL50803_7439	71770	X	X	
PP2A-2 Ser/Thr phosphatase PP2A-2 catalytic subunit	GL50803_5010	39777	X		11, 15, 18
PP2A Protein phosphatase 2A regulatory subunit B <sup>+</sup> , putative	GL50803_9894	101179	X		
Ser/Thr-protein phosphatase, catalytic subunit (metallophosphatase PP4-like)	GL50803_15214	35475	X		
Ser/Thr protein phosphatase, catalytic subunit (metallophosphatase PP2A-related)	GL50803_13524	44864	X	X	
PP2C Ser/Thr phosphatase, putative	GL50803_11740	39118	X		49
Hypothetical protein (TAP42 family protein)	GL50803_9058	50514	X		
Kinase, NEK	GL50803_101307	32501	X		
Kinase, NEK	GL50803_16824	90577	X		
Kinase, NEK <sup>4</sup>	GL50803_2483	57218	X	X	
Kinase, NEK	GL50803_26199	53055	X		
Kinase, NEK	GL50803_114937	85220	X		
Kinase, NEK	GL50803_114307	101473	X		
Kinase, NEK	GL50803_16967	154892	X		
Kinase, NEK	GL50803_87677	133217	X		
Kinase, NEK	GL50803_113553	71522	X		
Kinase, NEK	GL50803_16765	71410	X		
Kinase, NEK	GL50803_8856	61981	X	X	



1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

Kinase, NEK-frag	GL50803_7579	33846		X	
Kinase, NEK-frag	GL50803_5489	61331	X		
Kinase, NEK-frag	GL50803_4977	44937	X		
Kinase, NEK-frag	GL50803_102034	108210	X		
<b>Signal transduction</b>					
<b>14-3-3 protein</b>					
Kinase, STE (plant MAPKK/STE20-like) <sup>†</sup>	GL50803_6430	28577	X	X	
Kinase, CMGC GSK (Glycogen Synthase Kinase 3β/shaggy related protein kinase)	GL50803_22165	39536	X		18, 50
	GL50803_17625	39975	X		11, 51
Kinase, CMGC DYRK (DYRK1A/Yak1p/Ppk15p-like)	GL50803_137695	97249		X	52
Dual specificity phosphatase, catalytic (plant MKP1/Slingshot-like) <sup>†</sup>	GL50803_15112	78491	X		53
Kinase, CAMKK	GL50803_96363	300229	X		11, 54
Kinase, CAMK CAMKL (CIPK, CBL-interactin protein kinase homolog)	GL50803_16235	46098	X		
Kinase, CAMK CAMKL (SNF1/AMPK-like) <sup>†</sup>	GL50803_16034	49231	X		
AMPK, gamma-T non-catalytic subunit (SNF4-like)	GL50803_3414	39631	X		15, 18
Copine I (calcium-dependent, membrane binding protein)	GL50803_29490	29980	X		12
<b>Various</b>					
Variant Surface Protein with INR (VSP-186)	GL50803_14586	75531	X		
VSP	GL50803_112678	25337	X		
Hypothetical protein (VSP/high Cys membrane)	GL50803_114674	73024	X		
Leucine-rich repeat protein 1 virus receptor protein	GL50803_5795	85450	X		
CipB protein heat-shock protein HSP101	GL50803_17520	97764	X		
MYG1 protein	GL50803_10858	40860	X		
4-methyl-5-thiazole monophosphate biosynthesis enzyme (ThiI/DJ-1/PfpI superfamily) <sup>†</sup>	GL50803_9088	19639	X		
Nucleotide-binding protein 1 (NUPP/MRP subfamily of ATP-binding proteins)	GL50803_10969	36763	X		
Hybrid cluster protein (hydroxylamine reductase), lateral transfer candidate	GL50803_3042	62546	X	X	
CEP1 protein	GL50803_17120	58286		X	
<b>21.1 Proteins (unknown function)</b>					
Protein 21.1 with ANK and SMC domains	GL50803_113622	164122	X	X	
Protein 21.1 with ANK domains	GL50803_137703	56587	X		
Protein 21.1 with ANK and SMC domains	GL50803_17551	118715	X		
Protein 21.1 with ANK and SMC domains	GL50803_11107	89274	X		
Protein 21.1 with ANK domains	GL50803_6007	113222	X		
Protein 21.1 with ANK and SMC domains	GL50803_16532	92676	X		
Protein 21.1 with ANK domains	GL50803_24009	46684	X	X	
Protein 21.1 with ANK domains	GL50803_16915	84080	X		
Protein 21.1 with ANK domains	GL50803_16220	71121	X	X	
Protein 21.1 with ANK domains	GL50803_14480	114237	X		
Protein 21.1 with ANK domains	GL50803_7375	86846	X		
Protein 21.1 with ANK domains	GL50803_95192	124352	X		
Protein 21.1 with ANK domains	GL50803_11493	146965	X		
Protein 21.1 with ANK domains	GL50803_16533	102632	X		
Protein 21.1 with ANK domains	GL50803_7414	40625	X		
Protein 21.1 with ANK domains	GL50803_7373	169337	X	X	
Protein 21.1 with ANK and SMC domains	GL50803_93011	114347	X		
Protein 21.1 with ANK domains	GL50803_12139	75885	X		
Protein 21.1 with ANK and SMC domains	GL50803_4363	83939	X		
Protein 21.1 with ANK domains	GL50803_7679	78497	X		
Protein 21.1 with ANK domains	GL50803_13982	61372	X		
Protein 21.1 with ANK domains	GL50803_15587	27664	X		
Protein 21.1 with ANK domains <sup>†</sup>	GL50803_13436	21308	X		
<b>Unknown function</b>					
Hypothetical protein	GL50803_16699	188597	X		
Hypothetical protein	GL50803_16222	188519	X		
Hypothetical protein	GL50803_114921	224351	X	X	
Hypothetical protein	GL50803_4883	162837	X	X	
Hypothetical protein	GL50803_137693	79816	X		
Hypothetical protein	GL50803_17577	120226	X		
Hypothetical protein	GL50803_14003	100281	X		
Hypothetical protein	GL50803_7350	174590	X		
Hypothetical protein	GL50803_8586	96695	X		
Hypothetical protein	GL50803_16761	87014	X		
Hypothetical protein	GL50803_16504	97345	X		
Hypothetical protein	GL50803_10522	29761	X		
Hypothetical protein	GL50803_13330	26904	X		
Hypothetical protein	GL50803_15213	69974	X		
Hypothetical protein	GL50803_16999	138117	X		
Hypothetical protein	GL50803_13323	92275	X		
Hypothetical protein	GL50803_12105	69801	X		
Hypothetical protein	GL50803_8606	109433	X		
Hypothetical protein	GL50803_10568	40088	X		
Hypothetical protein	GL50803_6725	100211	X		
Hypothetical protein	GL50803_21063	75449	X		
Hypothetical protein	GL50803_32399	49183	X		
Hypothetical protein	GL50803_12770	25697	X		
Hypothetical protein	GL50803_9099	45649	X		
Hypothetical protein	GL50803_11034	24807	X		
Hypothetical protein	GL50803_7422	69314	X		
Hypothetical protein	GL50803_10882	33117	X		
Hypothetical protein	GL50803_23767	54462	X		
Hypothetical protein <sup>†</sup>	GL50803_7240	36199	X		
Hypothetical protein	GL50803_103540	32192	X		



1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

Hypothetical protein	GL50803_15110	20656	X		
Hypothetical protein	GL50803_10608	102993	X		
Hypothetical protein	GL50803_10232	73122	X		
Hypothetical protein	GL50803_34179	157786	X		
Hypothetical protein	GL50803_17362	189910	X		
Hypothetical protein	GL50803_14846	116965	X		
Hypothetical protein	GL50803_7188	115522	X		
Hypothetical protein	GL50803_15120	26644	X		
Hypothetical protein (with Leucine-rich domain)	GL50803_5543	168048	X		
Hypothetical protein contains (RING and CCH-type Zn-fingers)	GL50803_14241	153634	X		
Hypothetical protein (homology with armadillo repeat-containing protein)	GL50803_5784	120602	X		
Hypothetical protein (with WD40 repeats)	GL50803_15268	96102	X		
Hypothetical protein (with WD40 repeats)	GL50803_22573	242121	X		
Hypothetical protein (with ANK domain)	GL50803_11207	57965	X		
Hypothetical protein (with CBS domains)	GL50803_8692	37196		X	
Hypothetical protein	GL50803_4296	121302		X	
Hypothetical protein <sup>d</sup>	GL50803_8528	66851		X	
Hypothetical protein	GL50803_94463	53390		X	

<sup>a</sup>Protein proposed as potential novel drug targets (Morrison et al., 2007).

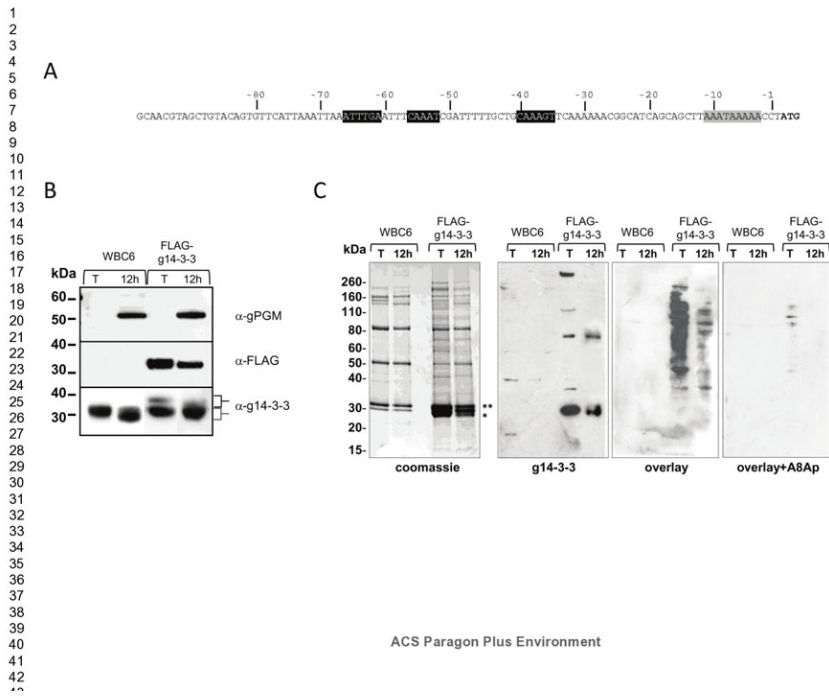
<sup>b</sup>Protein without 14-3-3 binding motifs based on Supplemental Table2.

<sup>c</sup>Identified protein domains (Pfam) are reported in bracket.

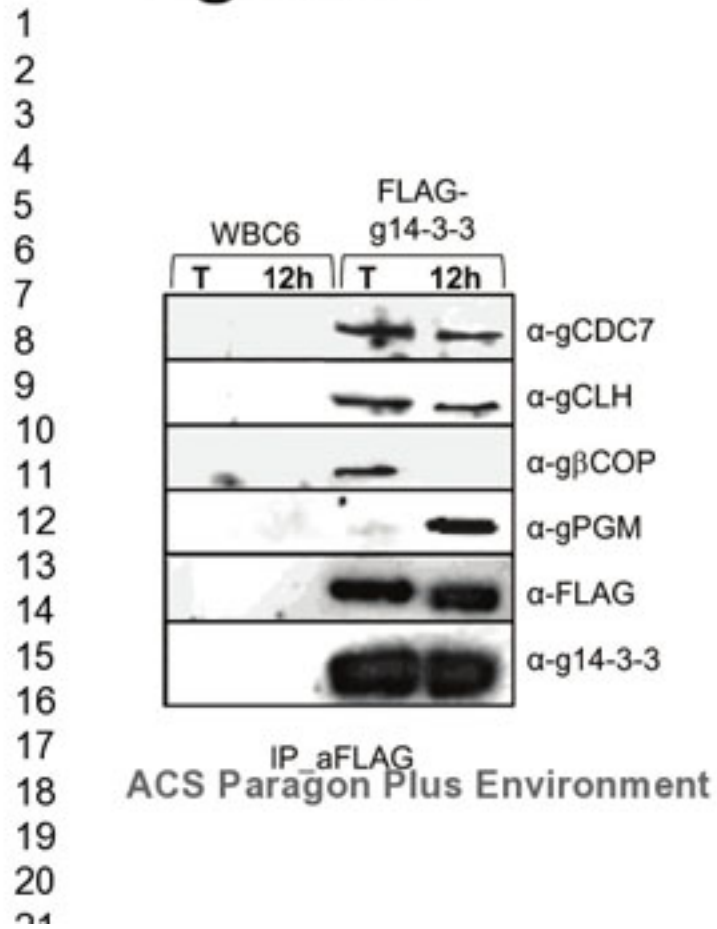
<sup>d</sup>Gray boxes indicated proteins identified in both trophozoite and encysting parasite samples, while black boxes indicate protein identified only in encysting parasite samples.

<sup>e</sup>Literature describing interactions of 14-3-3 with homologous proteins in other biological models in proteomic, protein-protein interaction or genetic studies.

Figure 1



## Figure 2



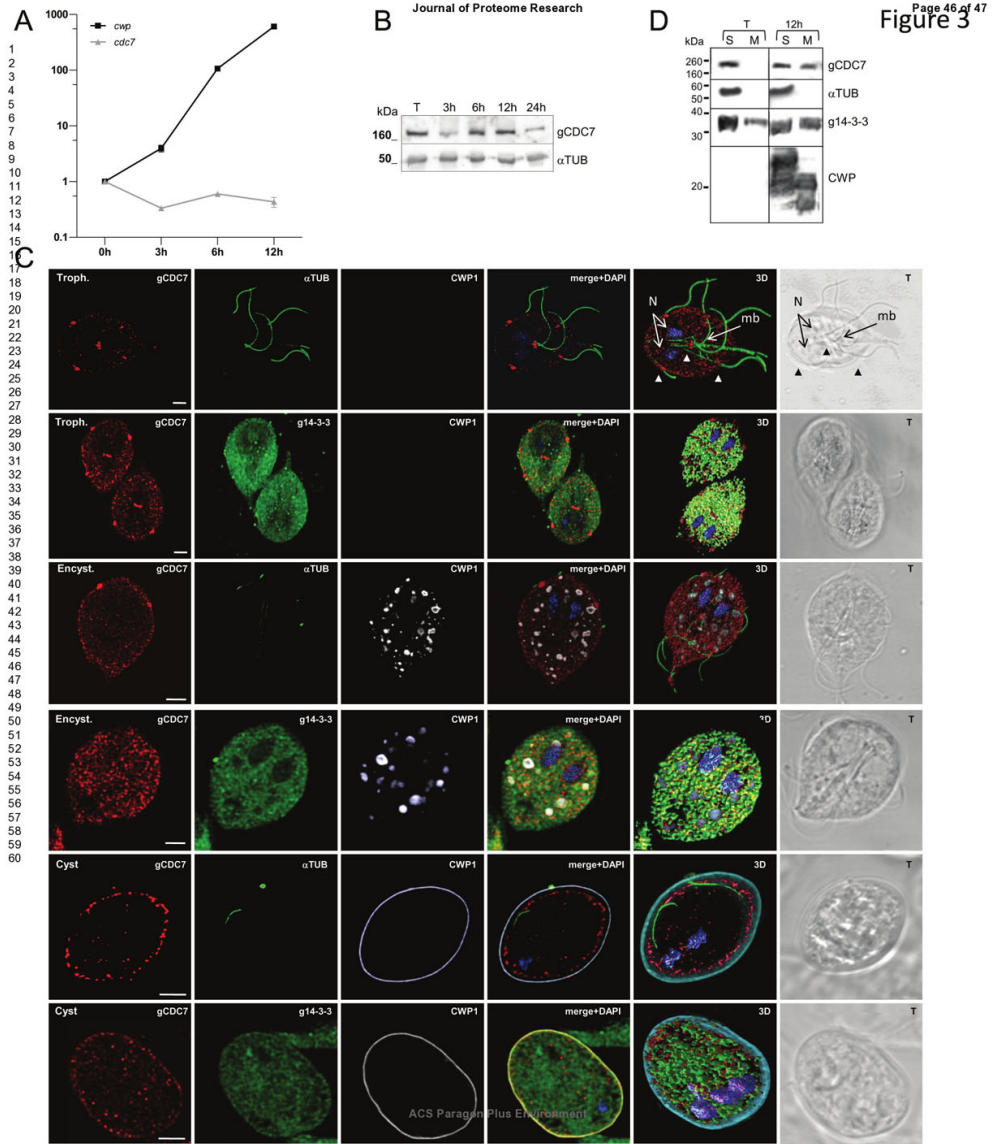


Figure 4

