Functional Sites in Proteins: analysis and annotation of five nuclear functional sites for the ELM resource

by

Morten Mattingsdal

A thesis submitted in the partial fulfillment of the requirements

for the degree *Candidatus scientiarum* in Molecular Biology

Department of Molecular Biology

University of Bergen

Bergen, Norway

May 2004

Acknowledgments

I wish to express my gratitude to my assistant supervisor Dr. Pål Puntervoll for having the courage for allowing me to start this work, despite little experience in bioinformatics and zero experience in Linux. Thanks also for reading the manuscript and making invaluable comments. Thanks to my supervisor Prof. Rein Aasland for having the patience to see me finish and always spawning new ideas. Thanks to both for allowing me to travel to the ELM consortium meetings and dine, drink and talk to the nice people behind the e-mail addresses, including Dr. Toby Gibson, Dr. Rune Linding, Christine Gemuend, Dr. Sophie Chabanis, Dr. Scott Cameron, Dr. David Martin, Prof. Manuela Helmer Citterich, Dr. Leszek Rychlewski and Dr. Francesca Diella.

I also wish to thank the people responsible for letting me lend a computer at the Computational Biology Unit at the University of Bergen. I suspect that this included Dr. Pål Puntervoll and Prof. Inge Jonassen.

Thanks to group-members Anja Ragvin og Katharina Tufteland for social events and understanding my frustrations. Thanks to Valve software for Counterstrike and Magnus Blø for playing it. Thanks to my father for purchasing a Commodore-64 fifteen years ago making Linux less intimidating. And many thanks to my mother for standing my whining and transferring pocket money when times were bad. Last but not least, I wish to thank myself for doing this.

"Endele dreit føle"

Morten Mattingsdal

Contents

Summary

Proteins are molecular tools which a cell expresses from specific genes, where the protein has one or several functions. An example of a protein function can be to detect a particular hormone, transporting itself to the nucleus and promoting transcription of other genes, as a response to the hormone signal. These three "actions" that the protein performs, can be narrowed down to regions inside the protein sequence. These regions are in this thesis termed functional units. In this example, one functional unit which recognizes the hormone, another unit which interacts with proteins that transport it to the nucleus and another unit which interacts with a DNA-bound protein, which results in transcription of target genes. Thus, several functional units defines the protein function. These functional units can be classified into different categories. One of these is functional sites.

Functional sites are small linear subsequences in proteins which can be related to biological function. Such a function may be modification sites like phosphorylation or protein-protein interactions sites like the LxCxE motif, which interacts with retinoblastoma proteins (Rbs). The problem with these sites is that they are so short, often not more then 3-5 amino acids in length. This implies an informatical problem when recognizing and predicting these short sites in protein sequences, which leads to a lot of hits and hence overprediction. The goal of the ELM project is to provide an Internet service for information retrieval and prediction of functional sites in proteins. The idea behind the ELM resource is to apply biological knowledge as context filters, and remove biological meaningless predictions, and thereby reducing the overprediction problem. For example, its meaningless to predict the LxCxE motif in extracellular proteins, since this functional site is only active inside the cell nucleus, where the Rbs are localized.

During this thesis five functional site have been annotated into the ELM resource. These functional sites are: LxCxE, SID, RxL, PxDLS and PxVxL.

Norsk oversettelse av sammendrag

Proteiner er molekylære verktøy som en celle utrykker fra et gitt gen der proteinet har en eller flere funksjoner. En biologisk kan være f.eks. å detektere et hormon i cytoplasma, transportere seg til kjernen for å så å promotere uttrykking av andre gener, som svar på signalet. Alle disse tre handlingene som proteinet utfører, kan snevres ned til regioner i protein sekvensen. Disse kalles her funksjonelle enheter. En funksjonell enhet som er i stand til å gjenkjenne hormonet, en annen som blir gjenkjent av andre proteiner som frakter det inn til kjernen og en siste enhet som er i stand til å binde til et annet protein som binder DNA. Dermed kan man si at flere funksjonelle enheter definerer protein funksjonen. Disse funkjonelle enhetene kan klassifieres i forskjellige kategorier. En av disse er funksjonelle seter.

Funksjonelle seter er små linære subsekvenser i proteiner som har en biologisk funksjon. Dette kan være modifiseringseter som fosforylering eller peptid ligander i protein-protein interaksjoner som for eksempel LxCxE motivet som binder til retinoblastoma proteiner (Rbs). Problemet med disse funksjonelle setene er at de er så korte, ofte ikke mer en 3-5 aminosyrer lange. Dette medfører et stort informatisk problem for å gjennkjenne og predikere disse setene i proteiner, fordi de vil ha mange treff og vil dermed overpredikere. Målsetningen med ELM prosjektet er å lage en database for funksjonelle seter i proteiner. Ideen bak ELM resursen er å implementere biologisk kunnskap som kontekst filtere, for å fjerne biologisk meningsløse prediksjoner, og dermed redusere problemet med overprediksjon. For eksempel gjør det ingen mening i å predikere LxCxE motivet i extracellulære proteiner, siden dette setet bare er aktivt i cellekjernen, der Rbs proteiner er lokalisert.

I løpet av denne hovedfagsoppgaven har fem funksjonelle seter blitt annotert i ELM resursen. Disse funksjonelle setene er: LxCxE, SID, RxL, PxDLS og PxVxL.

Definitions

Context: The surroundings or environment which makes up the prerequisites where an event takes place.

Region: A part of, or subsequence of a protein.

Protein Unit: A region inside a protein which can perform a function in a given context.

Secondary structure element: α- helix, β-sheet or neither

Globular Domain: A protein unit composed of more then one secondary structure element which has a spherical appearance and few conformations. Has a defined function *in vivo*.

Functional Site: A protein unit composed of one secondary structure element, where the functional important amino acids are arranged in a linear manner. Has a defined function *in vivo*.

Recognition Module: A globular domain which is able to act upon a functional site.

Non-Globular Region: A region inside a protein which has many conformations and no distinct structure.

Annotate: The act of manually storing data in a database.

ELM instance: A single protein which contains an experimentally verified functional site.

Siteseeing: Collecting ELM instances and annotating them into the ELM resource.

Homology: Two proteins are homologous if they share a common evolutionary ancestor.

Chapter 1

Introduction

This introduction discusses functional sites from a biological and applied bioinformatical perspective.

1.1 Functional units in proteins

There are many different types of proteins, ranging from structural proteins in the cytoskeleton to secreted enzymatic proteins. They all have one thing in common: they interact with the surrounding environment. Besides water, the surroundings consist most often of other molecules like: other proteins, DNA, hormones and other small organic molecules. When investigating a protein, and the following question is asked: "Why is my protein able to recognize another molecule?", the answer often reveals that a region of the protein is responsible. This is often a protein unit (see definition). Several protein units make up a modular protein, which sometimes allows a protein to act and respond to several events at the same time or in different situations.

Globular Domains. "Protein domain" is a phrase that is often used in different research areas with different meanings. For a structural biologist, a protein domain is a unit with a distinct structure and globular appearance. For an experimentalist, a protein domain could correspond to a protein region with a specific function. A globular domain (see definition) can be considered as multiple secondary structure elements that acts as a framework for orienting several specific residues in space, and thereby defin-

ing one or more biochemical environments which is responsible for a function, e.g. an active site in an enzyme. Here domain and globular domain are used synonymously[1].

Domains constitute the largest protein unit which are structured and have a globular shape, due to the presence of hydrophobic amino acids that play a central role in protein folding and stability [179]. Metal ligands, disulfide bonds and salt bridges are also major factors which influence domain folding and stability. The functionally important amino acids in a globular domain are often close to each other in space, but can be far apart in the protein sequence. Globular domains are composed of several secondary structure elements and can therefore arrange and orient several amino acid residues in specific positions that can perform a certain task, e.g. the Ser/His/Asp/Thr tetrad in some serine proteases [114]. Globular domains are sometimes referred to as a modules, reflecting genetic mobility and a particular evolutionary history [170]. A model for the genetic mobility of domains has been proposed: exon-shuffling [72]. This exon-shuffling mechanism is thought to play a central role in in the formation of novel compositions of modules (and novel proteins). This mechanism is also believed to explain the observation of multi-modular proteins in eukaryotes, especially in extracellular and nuclear proteins [176, 120].

Globular domains allow proteins to perform work in the cell, such as breakdown of nutrients, interaction with a specific DNA region or detecting signaling molecules or functional sites. Many globular domains are recognition modules, and some of these are shown in Fig 1.1.

Functional sites. Functional sites are short linear regions in proteins that have a specific function, and are often found in unstructured regions in proteins [180]. They are known by several different names: functional peptide, signature motif or active peptide. Here they are called functional sites. Functional sites can be divided into different classes based on their function: modification, ligand and cleavage sites[2]. Modification sites are recognized by proteins that covalently attaches or detaches molecules to the

[1] See Fig 1.2 for a schematic overview of different terms used here.

[2] The nomenclature in the ELM resource uses four classes: ligand (LIG), target (TRG), modification (MOD) and cleavage (CLV). The target and ligand functional sites are of the same category, but are separated due to functional differences. Where TRG are a LIG that determines sub-cellular localization of the host protein.

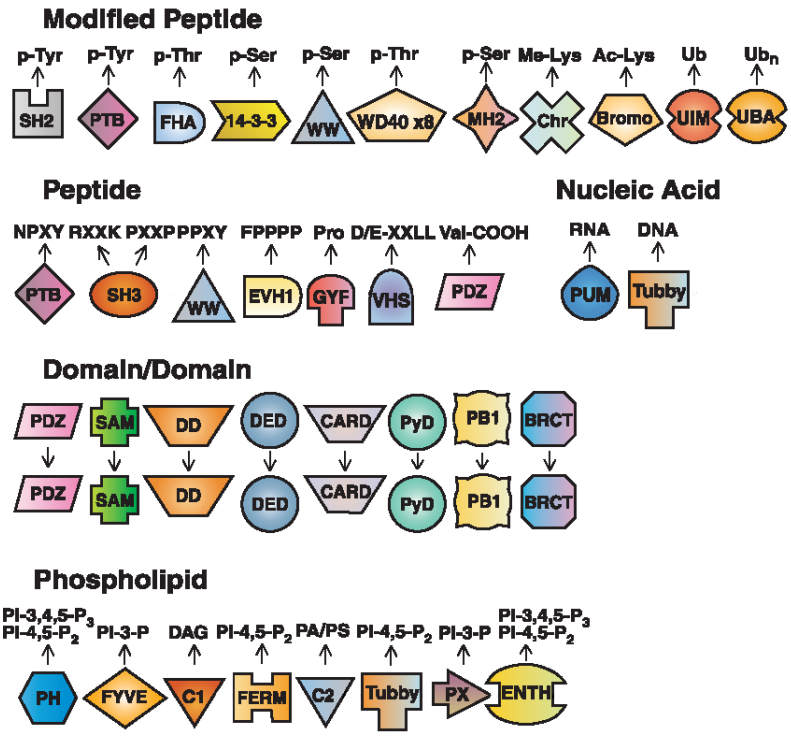


Figure 1.1: **Cartoon of some recognition modules involved in signal transduction.** By the definition in this thesis, the first two rows shows recognition modules. The others are not recognition modules since domains, nucleic acids and phospholipids are not functional sites. Adopted from [170].

sites, often with consequences for the target protein. One example is phosphorylation that regulates the activity of a globular domain. Cleavage sites are recognized by enzymes that break the peptide bond between amino acids. They have a central role in protein degradation and protein processing such as enzyme activation, turning latent precursor proteins into their biologically active enzymes. Ligand sites are functional sites that mediates interactions between proteins, and is a generic protein-protein interaction mechanism (Fig. 1.2). Ligand sites are able to mediate interactions with globular domains, used for example in the recruitment of co-repressors in transcriptional repression. One example of such a functional site is the LxCxE functional site which is found in several nuclear proteins which binds to the Retinoblastoma proteins and repress transcription [3] [49].

[3] The Retinoblastoma proteins and the LxCxE functional site is frequently used as an example throughout this thesis. See section 4.1 for a more detailed description about these proteins and the LxCxE functional site.

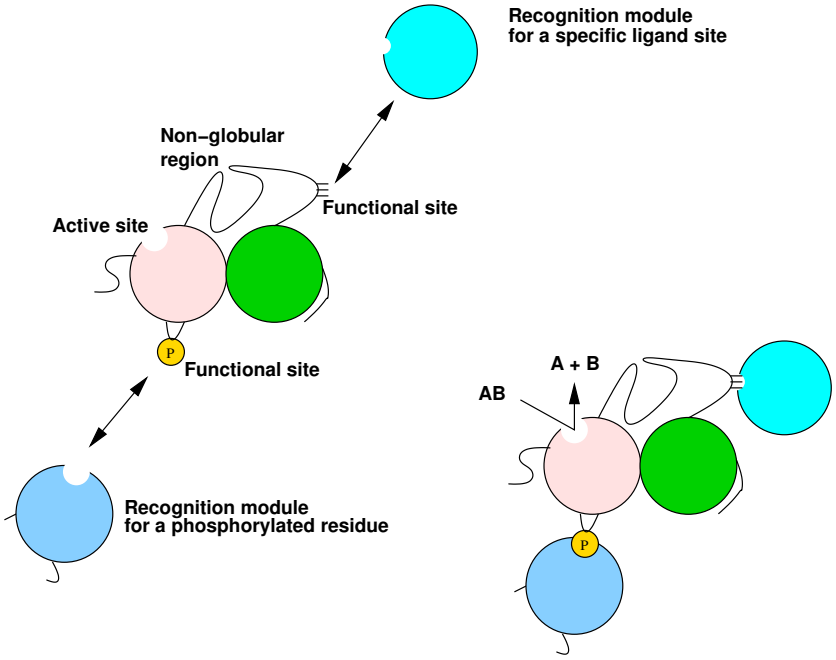


Figure 1.2: **Cartoon of different terms related to functional units.** The colored balls represent globular domains. Note that functional sites occur both outside and inside of globular domains.

All three classes of functional sites mentioned above, are actually ligands to their cognate recognition modules. After contact or proximity has been established, the site is cleaved, modified or retain the ligand status, depending on the recognition module. The defining feature of functional sites are that they are short, typical ranging from 4-10 amino acids, and that they are recognized and are acted upon by a globular domain.

Borderline functional sites. Globular domains and functional sites are two fundamentally different classes of protein units, which may represent two extremes. There may be additional views depending of the different definitions of globular domains and functional sites. The leucine-zipper, are by many, considered a domain. It is not globular, the functional important residues are organized in a linear manner in a quite long helix (~30 amino acids). For example seen in the basic helix-loop-helix leucine zippers of Mad-Max proteins (`SMART:HLH`)[4] [160]. The long length and the fact that leucine zippers are not recognized by a recognition module, but instead dimerizes, make the leucine zipper a borderline functional site. Functional sites are composed of only one

[4]Typewriter fonts are used to refer to annotations in the SMART, Pfam and ELM databases.

secondary structure element, and globular domains are composed of more then one secondary structure element. This implies that the basic helix-loop-helix leucine zipper may be considered a globular domain, but the leucine zipper, alone, a functional site.

Another borderline case are trans-membrane regions in proteins. These are composed of one secondary structure element and have the functionally important residues arranged in a linear manner, but it is not recognized by a recognition module. A final example of a borderline functional site is in the oligomerization domain of the p53 protein. See footnote 5.

Importance of context. A prerequisite for a protein unit to be functional, is that the unit must be in its appropriate molecular environment. This environment can be for example a sub-cellular compartment or in a preferable molecular region in a protein. A preferable molecular region can be in a non-globular region (see Fig 1.2), where a functional site is exposed and available to its cognate recognition module. It would be surprising if a region in a protein carrying the LxCxE motif described above, is functional in an extracellular environment. The functional site is taken out of context and inserted into an alien environment where it is unable to recognize or be recognized e.g. other protein units that it has co-evolved with. *All protein units require their appropriate context to be functional.*

1.1.1 Functional units in the Src and p53 proteins

The Src protein. A good example of how different functional units define different functional aspects of a protein function is the membrane bound tyrosine kinase Src, which is involved in signal transduction [143]. Tyrosine phosphorylation is a central mechanism in the regulation of a variety of biological processes such as cell proliferation, migration, differentiation and survival. Several families of receptor and non-receptor tyrosine kinases control these events by catalyzing the transfer of phosphate from ATP to a tyrosine residue of specific cell protein targets. The Src tyrosine kinase is composed of several protein units, as shown in Fig 1.3. When the SH2 ligand (functional site and “D” in figure) is phosphorylated, it binds to the SH2 domain,

leading to inactivation and inhibition of the Src tyrosine kinase [169]. A second functional site ("B" in figure) resides in a linker between the SH2 and SH3 domains. This site acts as a ligand to to the SH3 domain in the inactive form of the Src protein [92]. Hence, three functional sites are involved in the regulation and inactivation of the protein. Phosphorylation of the tyrosine residue, the phosphorylated tyrosine acts as a ligand, and the SH3 ligand in the so-called linker region between the SH2 and kinase domains. In addition, the protein has an N-terminal myristoylation site ("A" in figure), which is recognized for attachment of a myristoyl molecule anchors the protein to the membrane [140].

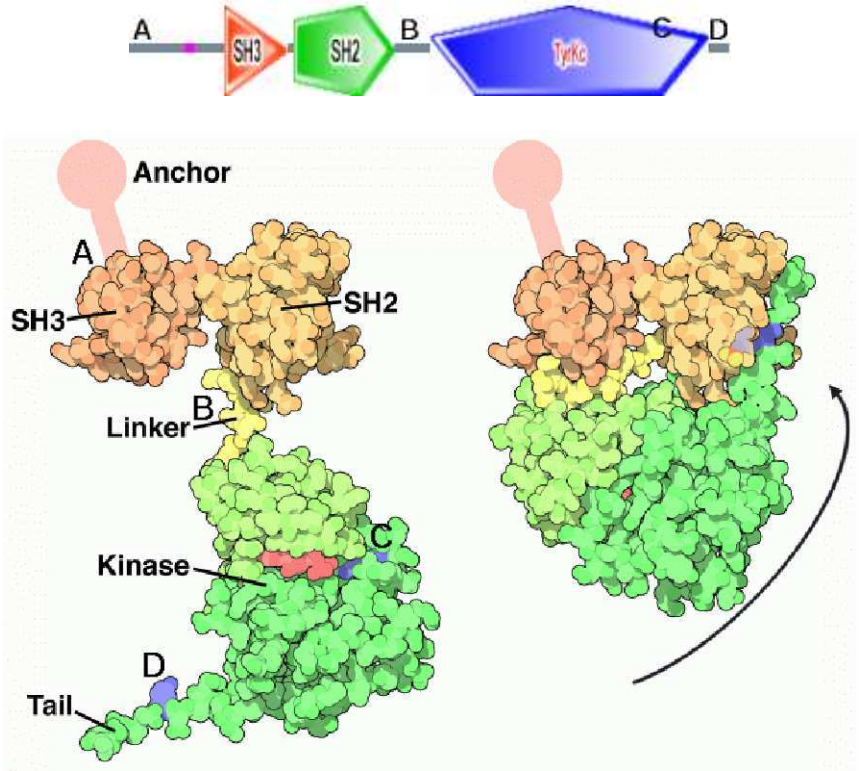


Figure 1.3: **Functional units in the Src tyrosine kinase.** Top, a cartoon showing the functional units in the Src tyrosine kinase. The domains predicted by SMART and four functional sites. A: myristoylation site, B: SH3 ligand, C: Auto phosphorylation site and D: CSK phosphorylation site & SH2 ligand. Below, active form (left) and its inactive form (right). Below picture adopted and modified from "Molecule of the Month", Protein Data Bank.

The p53 protein. Another example of how different functional units act together is in the human p53 protein [47]. The p53 protein is probably one of the most studied

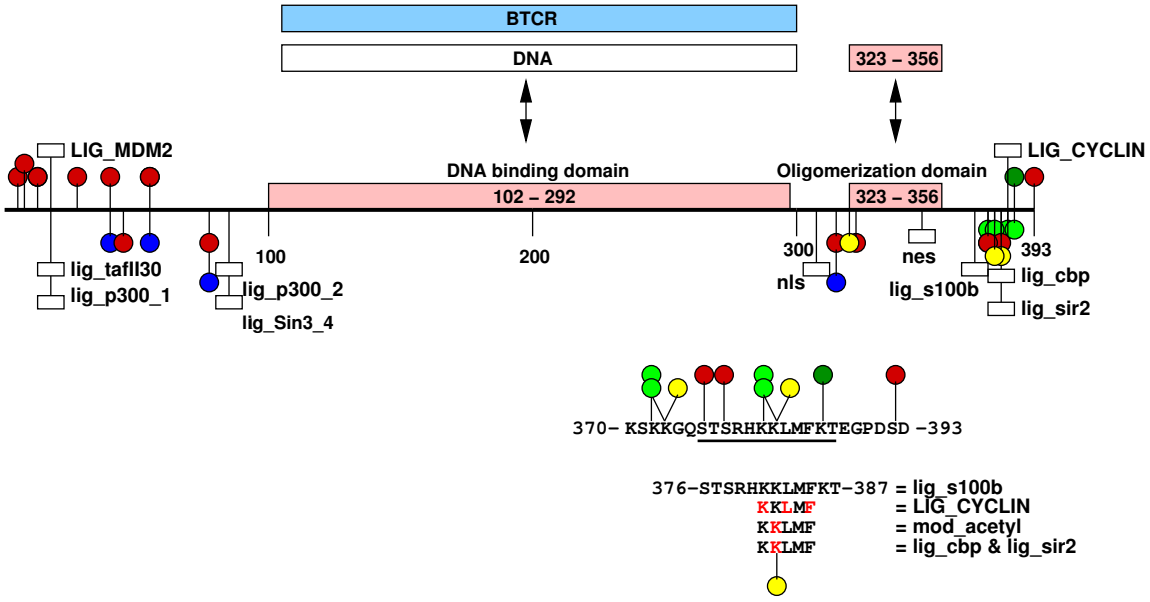
proteins, due to the close relationship between a dysfunctional p53 and cancer [91]. The concentration of p53 in normal cells is very low, but is increased as a response to cellular stress such as hypoxia (lack of oxygen) or DNA damage e.g. by radiation [118]. The p53 protein consists of a large number of functional sites and two domains (Fig. 1.4). The major function in the central DNA binding domain is to recognize specific genomic sites with the consensus 5'-PuPuPuC(A/T)(T/A)GPyPyPy-3 [58]. The central DNA binding domain is also able to mediate a protein-protein interaction with the BTCR domain found in the 53BP1 protein [53]. p53 binds as a tetramer, which the oligomerization domain[5] is responsible for, and promotes transcription of genes, via requirement of the basal transcription machinery through the functional site lig_tafII30 (Fig 1.4). Gene products of p53 promoted transcription include Bax, Bak and PUMA, which result in apoptosis and cell death [78]. Since p53 is a suicide protein, one would expect that expression and activation of the p53 protein needs to be highly regulated.

The need for p53 to be regulated is reflected by the high number of different modification sites in the protein (currently 25 sites were registered in this thesis). Half of the functional sites involved in protein-protein interactions in p53 are directly associated with an enzymatic process. The role of these sites may be to increase the time of proximity between an enzyme and p53, thus allowing an enzyme to discriminate between proteins and recognize its substrates.

1.1.2 Functional sites in the cell cycle and medicine

Functional sites are abundant in the cell cycle. Functional sites are important in protein function, as seen with the Src and p53 proteins (Fig 1.3 and 1.4). But what role do functional sites play in a larger context such as a pathway or a cellular process? Fig 1.5 shows a scheme of proteins involved in the different stages of the cell cycle. Most protein complexes in the figure are held together by domain-domain interactions, like that of cyclin/CDK, and most regulatory events involves functional sites. Phosphorylation of different proteins are very important in regulating the different stages of the cell cycle, and are the most frequent functional sites in Fig 1.5. One example is the

[5]Here the oligomerization domain is called a "domain", but it is in fact a borderline functional site where a single helix mediates the oligomer formation [146].



Functional site class	Functional site	Function	Ref.
Modification	phosphorylation	increased DNA affinity	[9]
	acetylation	increased DNA affinity	[96]
	ubiquitination	nuclear export and breakdown	[133]
	MOD_SUMO	unclear	[156]
	isomerization	increased DNA affinity	[235]
Ligand	LIG_MDM2	ubiquitination, degradation	[96]
	lig_tafII30	activation	[98]
	lig_p300_1 & _2	acetylation, activation	[54]
	lig_cbp	acetylation, activation	[155]
	lig_s100b	protection from modifications	[185]
	lig_sir2	deacetylation, repression	[17]
	lig_Sin3_4	deacetylation, repression	[236]
	LIG_CYCLIN	increased phosphorylation	[136]
	nls	nuclear import	[126]
	nes	nuclear export	[203]

Figure 1.4: **Functional units in the p53 protein.** The p53 protein have a high number of functional sites, ~30, and two globular domains. A central DNA binding domain and an oligomerization domain (both shown in pink). White squares are functional sites which acts as ligands in protein-protein interactions. Circles symbolizes modification sites and are colored as followed: Blue= isomerization, Red= phosphorylation, Yellow= acetylation, Green= ubiquitination and Dark Green= Sumoylation. Functional sites above the string and written in upper cases, are predicted by the ELM resource.

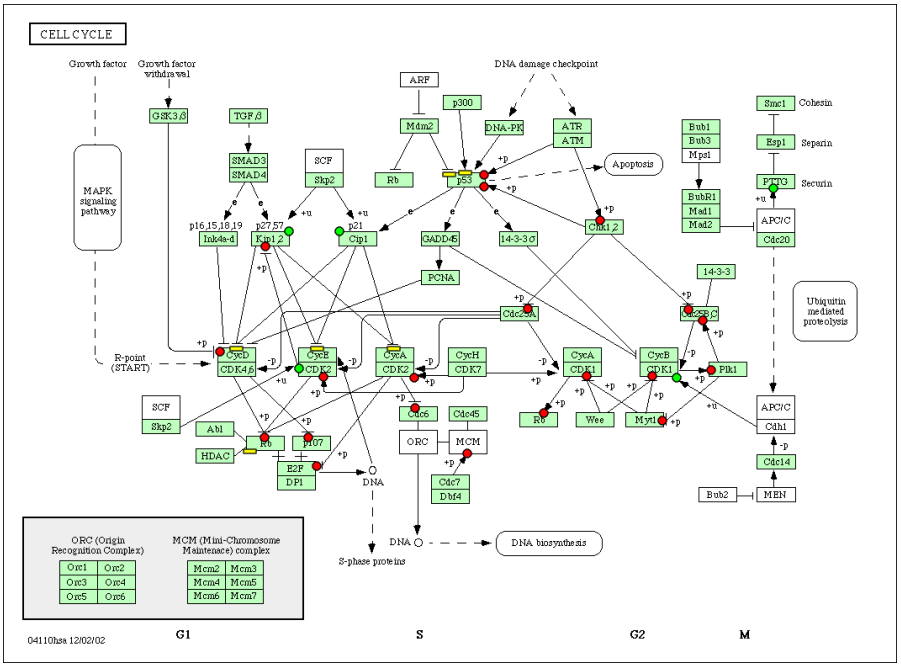


Figure 1.5: **Functional sites in the cell cycle.** Red circles represent phosphorylation, green are ubiquitination sites and yellow boxes are ligand sites. Picture adopted from the KEGG database [105].

Rb protein in the lower left of Fig 1.5. Rb binds to the E2F-DP1 hetrodimer through a domain-domain interaction [122]. Simultaneously Rb recruits repressors, like HDAC, though a functional site (see also section 4.1) to repress transcription of genes which are needed in the S phase of the cell cycle. To allow progression into the S phase, Rb becomes phosphorylated (at several functional sites) by cyclin/CDK complexes (see also section 4.3), to relieve repression and promote transcription of S-phase genes. One could state that functional sites are required to achieve the dynamic and regulatory complexity seen in Fig 1.5. The cell cycle is one of numerous cellular processes. Presumably functional sites are abundant in most cellular processes, especially in processes where post-translational modifications are involved in e.g. cell signaling and signal transduction [171].

Recognition module-functional site interaction in medicine. Recognition modules have been investigated as targets of drugs, to inhibit an interaction between a recognition module and the corresponding functional site. Cancer cells are uncontrolled replicating cells and disturbing or inhibiting proteins important in the cell cycle

are targets for several drugs.

An example of a recognition module which is a target for drugs, is a cyclin domain found in cyclin proteins. This recognition module is able to interact with peptides and proteins which have the RxL functional site. This interaction is important in the recognition and subsequent phosphorylation of cyclin/Cyclin Dependent Kinase (CDK) substrates. Since cyclin/CDK dependent phosphorylation is an important process in regulating the cell cycle (Fig. 1.5), inhibiting the interaction between cyclins and RxL containing proteins, can lead to the disturbance and death of replicating cancer cells [149].

Another example of targeting a recognition module, is the p53 binding domain of MDM2. MDM2 is a ubiquitin ligase which targets the p53 protein (see section 1.1.1) for the attachment of the ubiquitin protein, leading to ubiquitin-dependent degradation by the proteasome. A functional site in the p53 protein, the Fxx motif [211], interacts with the recognition module "p53 binding domain" of MDM2. MDM2 is a negative regulator of p53, and inhibiting the interaction between these two proteins may increase the concentration of p53, leading to a possible proliferation arrest of cancer cells [43].

Peptides and small organic compounds that mimic peptides and target recognition modules are subject to intensive research to have potential applications as drugs [12].

1.2 Predicting functional units

The traditional way of predicting protein function, is to compare the whole sequence of a protein in question with proteins of known functions, and using sequence similarity to infer homology (see definition) and homology to infer function. The Basic Alignment Search Tool or BLAST [6], is the most frequently used sequence search method. BLAST compares a query sequence with sequences in a database, and scores them based on similarity. Domains, being the largest of the protein units, influence scoring of BLAST significantly, while other shorter units like functional sites have a smaller contribution on the final scoring, and are thereby often not detected and ignored.

Another approach in predicting protein function, is to identify the functional units

which may reside in a protein sequence. Several bioinformatical approaches are available which uses scoring or probability matrices for prediction of functional units [76]. These include PSSM (Position Specific Scoring Matrices) [86] and HMM (Hidden Markow Models) [56]. PSSM and profile HMMs are often applied to globular domains [28] and HMM based methods are used in predicting trans-membrane units [198]. These methods are, in many cases, not applicable to functional sites since they are to short and often statistically insignificant.

Regular expressions can be applied to the identification of conserved regions in proteins and identification of domains and functional sites. A pattern or a regular expression[6] is a sequence or a set of rules that is matched against a string of text. A match to a regular expression is binary; either the regular expression matches or it does not. This makes it difficult to score a match against a regular expression. Functional sites are intrinsically short and overprediction have been and still are a problem. When a method overpredicts, usually means that the method generates many hits, and only a fraction of the hits correspond to true positives. Although challenging, countermeasures can be applied to reduce the number of hits and hence reduce the overprediction problem.

Web Resources for prediction of domains. Several online resources are dedicated to prediction of domains in proteins and some of these are listed in table 1.1 [7]. Most of these databases use profile based methods like HMM or PSSM and focus on long conserved regions in proteins. Smaller regions, like functional sites, are more troublesome and several strategies are used for predicting these.

Web resources for prediction of functional sites. Numerous databases exist for prediction of functional sites (table 1.2), focusing on particular themes. The most generic resource have been Prosite. But since functional sites are so short and difficult to predict by sequence detection methods alone, the resource has emphasized domain annotation instead of functional sites [180]. Presumably the main reason for this is

[6] In this thesis patterns are used to refer to the Prosite language, and regular expressions refer to the POSIX regular expression syntax, see section 3.4

[7] This is the main focus of the resources. Additional predictions are also performed e.g. trans-membrane and low complexity regions.

Table 1.1: **Resources for prediction of protein families and domains.**

Database	Short description	Ref.
SMART	Domains	[125]
Pfam A	Domains and protein families	[27]
Prosite	Domains, protein families and functional sites	[23]
ProDom	Generated from Swiss-Prot and TrEMBL	[45]
TIGRFAMs	Protein families	[80]
PRINTS	Protein families	[16]
InterPro	Unification of all databases above	[10]

Table 1.2: **Specialized resources for prediction of functional sites.**

Database	Short description	Ref.
PredictNLS	Nuclear localization signals	[159]
TargetP	Sub-cellular location	[61]
PSORT	Localization sites	[161]
SignalP	Cleavage sites and signal peptides	[164]
NetPhos	Phosphorylation sites	[30]
The Sulfinator	Tyrosine sulfation	[152]
NetOGlyc	GalNAc O-glycosylation	[83]
NMT	N-terminal N-myristoylation	[145]
big-PI	GPI Modification Sites	[57]
iSPOT	SH2, 14-3-3 or PDZ binding motifs	[33]
ScanSite	SH2, SH3, 14-3-3 and PDZ	[165]
Prosite	Functional sites and domains	[23]

that functional sites are so short and difficult to predict in protein sequences. Prosite have the comment “Warning: pattern with a high probability of occurrence”, to notify the user that the prediction is not to be fully trusted.

1.2.1 The ELM resource

As seen in the Src and p53 proteins, functional sites are an important aspect of how proteins function and behave in the cell. Functional site are intrinsically short and detecting these in protein sequences are prone to overprediction. The reason for this overprediction is mainly due to the detection method, but also the false assumption that all protein sequences are candidates to contain any given functional site. Functional sites are only active in their appropriate cellular or molecular context. By including the contextual information associated for each functional site, meaningless predictions can be removed and the overprediction problem can be reduced. This is the main motivation for the ELM resource. Take the following example: the functional site LxCxE, see section 4.1. This functional site is only active in the nucleus in higher eukaryotes. This functional site can be represented as the regular expression `[LI].C.E`[8] which has 6127 matches in Swiss-Prot and 484 matches in relevant proteins which fulfill contextual requirements (see table 5.3). Applying contextual information for this functional site leads to a 12 fold reduction of candidate proteins. The Eukaryotic Linear Motif (ELM) resource aims at being a resource for prediction functional sites in proteins, where biologists can perform relevant predictions of putative functional sites in protein sequences and browse the collection of functional sites that are collected and annotated. Se Figure 1.6 for flow scheme of prediction approach applied to the ELM resource.

The ELM resource has currently three filters: taxonomic range (NCBI taxonomic ID), sub-cellular localization (GO terms) and globular clash filter (SMART predictions). Additional filters are awaiting implementation (see section 5.5).

[8]Regular expressions are written in typewriter fonts to distinguish them from the name of the functional sites which often are consensus sequences. Name: LxCxE, regular expression: `[LI].C.E`

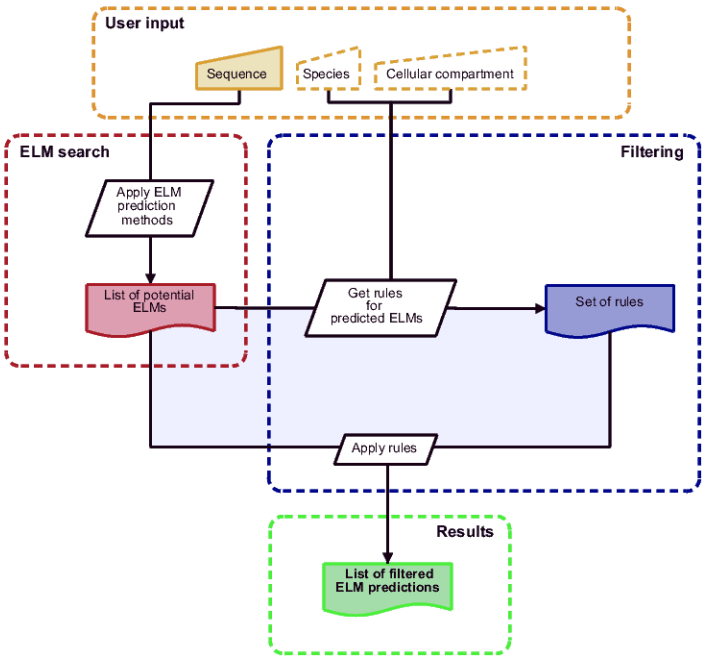


Figure 1.6: **Flow scheme of prediction strategy in the ELM resource.** Top: a user submits a protein sequence and taxonomic/cellular compartment ranges. Middle: the resource then searches for annotated regular expressions hits in the submitted sequence. After this, the rules of each regular expression are applied, and hits with a regular expression which does not meet the requirements of rules are removed. Bottom: Final results. Adopted from poster by Puntervoll & Mattingsdal.

Chapter 2

Background and Aims

2.1 Background

This work started in January 2002 when the ELM resource was in its infancy. The infrastructure of the database was ready, but lacking any annotated information on functional sites and therefore not able to do any predictions. The first version of the ELM server was launched in November 2002, and contained 30 annotated functional sites. During the next 16 months 61 additional functional sites were added, bringing the total number of annotated functional sites up to 91 (April, 2004). This thesis covers 5 of these 91 functional sites.

The ELM consortium consists of members that have various areas of biological interest and skills like: database design, software development, siteseeing (see definition) and development of filters. The ELM consortium involves five European academic institutions and one biotechnology company: European Molecular Biology Laboratory (European Molecular Biology Laboratory (EMBL), Germany), University of Bergen (Norway), University of Dundee (Scotland), University of Rome - Tor Vergata (Italy), BioInfo.PL (Poland) and CellZome (Germany). The focus at the University of Bergen are protein functions in the nucleus which involve functional sites.

2.2 Aims of this thesis

The main objective of this work was to collect data and annotate several nuclear functional sites for the ELM resource: aim **i)**, **ii**) and **iii)**.

i) Collect as many as possible experimentally verified proteins from the literature which describes the selected functional sites.

ii) Define regular expressions based upon the verified proteins for each functional site.

iii) Define context rules which are associated with each functional site.

iv) Register new candidate functional sites for the ELM consortium, for future analysis.

Chapter 3

Methods and data resources

This section is quite unconventional compared to traditional molecular biology theses at the University of Bergen. This is due to the lack of any experimentally derived data by the author, but instead the application of experimental data provided by the scientific literature. There is no exact method for manually collecting, analyzing and annotating data. Several approaches can be used, and the following chapter describes one possible approach.

3.1 Online tools and databases

The kind of work presented here, depends on the availability of other informatical tools and services provided by the scientific community. Following are world wide web tools and databases which are the major sources used in gathering different types of information regarding functional sites.

UniProt. UniProt or the Universal Protein Resource [11] is a comprehensive collection of protein sequence information consisting of the protein sequence databases: Swiss-Prot, PIR and TrEMBL [22, 26, 21]. UniProt was used as the source for retrieving protein sequences, as described in 3.3.1.3.

PDB. The Research Collaboratory for Structural Bioinformatics (RCSB) is a consortium that provide access to the protein data bank, or PDB, which is the main source

Table 3.1: **Online www tools and resources used in the siteseeing process.**

	Name	Purpose	URL
Databases			
	UniProt	retrieving protein sequences	http://www.ebi.uniprot.org/
	PDB	retrieving protein structures	http://www.rcsb.org/pdb/
	GO	retrieval of GO terms	http://www.geneontology.org/
	MEDLINE	database of abstracts from journals	http://medline.cos.com/
Tools			
	NCBI tax. b.	retrieving taxonomic ID numbers	http://www.ncbi.nlm.nih.gov/Taxonomy
	PubMed	searching MEDLINE	http://www.ncbi.nlm.nih.gov/entrez/
	Google	collecting information	http://www.google.com
	SRS	searching in biological databases	http://srs6.ebi.ac.uk/
	BLAST	taxonomic determination	http://www.expasy.org/tools/blast/
	ELM	functional site predictions	http://elm.eu.org/
	ScanProsite	retrieving protein ID	http://us.expasy.org/tools/scanprosite/
	Weblogo	conservation in an alignment	http://weblogo.berkeley.edu/
	SMART	prediction of globular domains	http://smart.embl-heidelberg.de/

for 3D structures of biological macromolecules [29]. The PDB databank was used to retrieve and study the structures of the functional sites, as described in 3.3.1.6.

GO. Gene Ontology is a database consisting of a set of defined biological terms describing how gene products behave in a cellular context [14]. GO provides biologists with a controlled set of terms that simplifies annotation of biological data and computational analysis of databases annotated with GO terms. GO terms were assigned to each functional site as explained in 3.3.2.2.

PubMed/MEDLINE. PubMed/MEDLINE is the main source for publications regarding scientific biological literature registered at the U. S. National Library of Medicine.

NCBI taxonomy browser. The NCBI taxonomic browser is a a tool at the U.S. National Center for Biotechnology Information [225]. The NCBI taxonomic browser was used to retrieve the NCBI taxonomic identification numbers for the taxonomic annotation of each functional site.

Google. Google is an Internet search engine which was very useful in collecting information not published in PubMed, like information stored in other biological databases like FlyBase [15] or in the personal homepages of scientists.

SRS. SRS or the Sequence Retrieval System is a virtual library which allows queries in different databases. SRS offers a user specific searches where a query may be a protein name derived from a publication. For protein queries, the database UniProt was used.

BLAST. BLAST is an algorithm developed as a method for rapid sequence comparison, and several websites offer this service, including; ExPASy, EMBL and NCBI [7]. BLAST searches were used in determining the taxonomic distribution of each functional site, see 3.3.2.1.

ELM. The ELM server was primarily used here to investigate if a functional site resides inside a domain sequence. In addition the ELM server was used to predict functional sites inside putative true positive proteins and browse other annotated functional sites [180].

ScanProsite. ScanProsite is a tool which allow searches with a user defined prosite pattern or a annotated prosite pattern in various databases. The tool was useful in identifying a protein, where only a subsequence is known, see 3.3.1.3. ScanProsite is a “child” of Prosite [23].

Weblogo. Weblogo is a web based application for analyzing an alignment by generating a picture where the size of the one letter code letters correspond to their frequencies in an alignment [190].

SMART. SMART is a tool for prediction of domains in a protein sequence [125]. SMART was used here to explore if the functional site occur inside a domain sequence and if some domains can be related to a functional site as an co-occurrence, see 3.3.2.3.

Table 3.2: **Offline tools used in the siteseeing process.**

Name	Topic
Vim	visualization of regular expression hits
grep	counting hits performed by elmfetch
ClustalX	making and coloring of alignments
elmfetch	extract subsequences from a sequence file
getseq	retrieval of sequences
PyMOL	visualization of protein structures
VMD	visualization of protein structures

3.2 Offline tools

In table 3.3, are a description of small programs and tools used locally in siteseeing. A Linux platform were used to run the following tools.

Vim. Vim is a Unix text editor which is used to create and edit text files. Vim was used to navigate and visualize regular expression hits in protein sequence files, see section 3.3.1.5.

Grep. Grep is a Unix application which is used to search an input file for lines containing a match to a specified pattern. Grep was here used to count the number of hits performed by elmfetch, see 3.3.1.8.

ClustalX. ClustalX is a graphical interface program for the ClustalW multiple sequence alignment program [207, 208]. ClustalX was primarily used here to color and make a picture of an alignment of functional sites. In addition ClustalX was used to make an alignment of homologous proteins. Often with default settings, but identity matrix, at multiple alignment options, showed to be particular useful when aligning a functional site in homologs proteins.

elmfetch. elmfetch is a script, written by Dr. Pål Puntervoll, which scans a protein database, in FASTA format, for a user defined regular expression. The results are displayed in a file describing the position of a hit in a sequence. Optional parameter

is including flanking regions of the hit. All searches in local protein databases with patterns were performed with elmfetch. See 3.3.1.5 and 3.3.1.8.

getseq. getseq is a script written by myself and with very good help from BioPython cookbook [101] and a Python tutorial [214]. The program queries ExPASy online for a FASTA sequence and prints it on the screen. The user must provide an accession number. See appendix for code.

PyMOL. PyMOL is an modeling system, which include rendering capabilities for making pictures of protein structures. URL: http://pymol.sourceforge.net/.

VMD. VMD is a molecular visualization program for displaying, animating, and analyzing protein structures. URL: http://www.ks.uiuc.edu/Research/vmd/

3.3 Siteseeing

The following procedure is the approach used in annotation of all five functional sites described here.

3.3.1 Going from individual papers to a regular expression

3.3.1.1 Identification of a functional site in the literature

A large amount of time was used in the identification of scientific publications describing an interaction[1] which involves one of the five functional sites in this thesis. Keywords describing the interaction, the name of the protein harboring the recognition module or the name of the biological process were often used as search words in PubMed/MEDLINE. Reviews or structural publications did almost always provide a good start for further collection of more experimental publications. If several publications described the same interaction, the most detailed publication was kept. Each publication was printed out for a more thorough examination of the experimental data used and which region or subsequence was responsible for the interaction. The PubMed identification number (PMID) was stored for future annotation in the ELM database.

[1] All functional sites in this thesis are ligands in a protein-protein interaction.

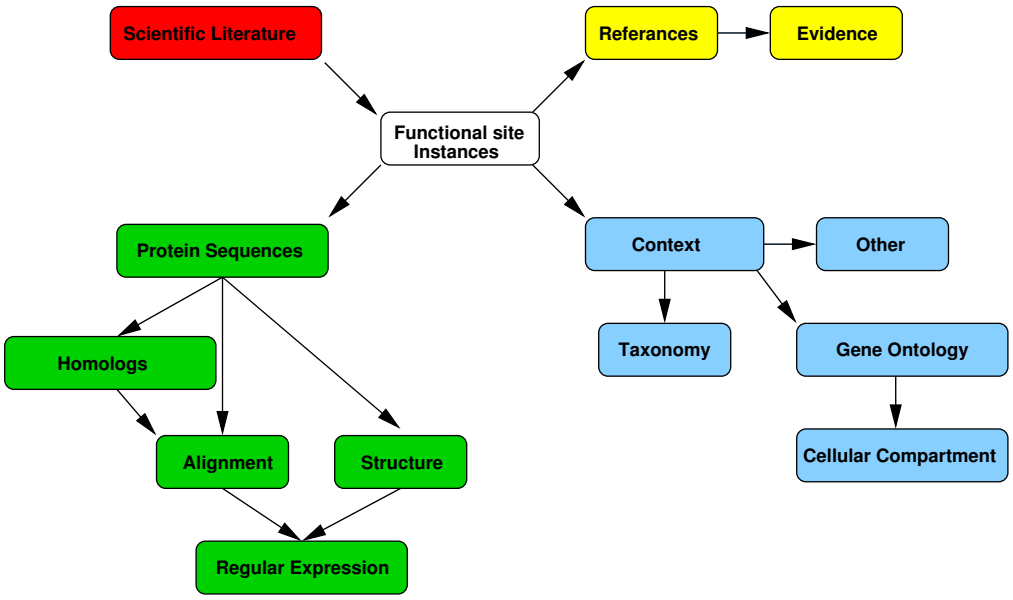


Figure 3.1: **Siteseeing.** Flow scheme over the siteseeing process used in this thesis. The results of siteseeing are; regular expression that describes the functional site, context rules that are associated with the regular expression and annotated evidence for each ELM instance.

3.3.1.2 Determination of protein name

The collected publications from 3.3.1.1, were examined for identification of the name of the protein containing the functional site. The names were stored separately for future sequence retrieval.

3.3.1.3 Retrieving protein sequences

The name of the proteins containing a functional site, from section 3.3.1.2, were submitted in SRS to query and subsequent retrieval of protein sequences from the database UniProt. The script “getseq” was also used for retrieval of protein sequences. The retrieved protein sequences were stored in FASTA format [172]. FASTA format was used since it is the input format for the script “elmfetch”. All protein sequences containing the same functional site were stored in the same file, e.g. the LxCxE functional site (section 4.1) have been identified in 25 proteins, hence 25 protein sequences in the same file. This file was subjected to analysis by elmfetch and manual inspection in vim. In some cases it was difficult to retrieve the protein sequence. Presumably the name of the protein in the publication differed from the name annotated in the

database, or the protein have not been annotated at all. If a publication supplied a subsequence, ScanProsite was sometimes used in the identification of the ID or accession number of the protein.

3.3.1.4 Expand the dataset by homology

Homologs were only included if a publication suggested or showed functionality of homologs. BLAST was used to explore potential functional homologs, but often revealed an altered functional site which differed from the experimental derived sequences. BLAST was not used in a consistent manner mainly due to the following reasons: 1) collecting homologs from all verified proteins containing a functional site in this thesis, a total of 97 proteins, is a considerable workload. 2) difficulty in determining functionality e.g. if the functional site and cognate recognition module has diverged in taxonomic lineages.

3.3.1.5 Extracting subsequences and building the alignment

The files containing the proteins in FASTA format, from section 3.3.1.3, were used to extract subsequences containing the functional site in question. A prerequisite for doing this is that a preliminary regular expression is available. A preliminary regular expression is the immature regular expression which can be derived from the consensus sequence or a proposed pattern in the literature. Manual inspection of each protein sequence was necessary for further development of the regular expression. The Unix editor, vim, was used for manual inspection of the FASTA formated protein sequence file. Vim allows regular expression searches and corresponding hits are colored. This approach was used to manually compare regular expression hits with reported regions in the collected publications, and subsequent refining of the regular expression. When a regular expression detects every verified functional site in the FASTA formated sequence file, the script elmfetch was used for extraction of the subsequences. The output from elmfetch from each case can be seen in the alignments in Chapter 4, where the output from elmfetch were visualized in ClustalX.

3.3.1.6 Structural information

Structural data were available for all five functional sites analyzed here. The name of the proteins containing the recognition modules were used as queries in the PDB database. The VMD tool was used to explore the structure, and PyMOL was used to make pictures of the functional sites. Structural information was used to refine the regular expressions, especially in the determination of variable positions. For example, if the structure reveals that a side chain of an amino acid is facing the solvent and not the recognition module, this position was often given an "x" (meaning: any amino acid is allowed to occupy this position). The importance of different positions are well described in the corresponding publication of the structures.

3.3.1.7 Extraction of the regular expression

The regular expressions were continuously changed and refined as more information were gathered. A good alignment of the subsequences is a very good start for the development of good regular expression. After the generation of the alignment, from section 3.3.1.5, a Weblogo was made for easy visualization of important residues. The final regular expression should be able to detect all experimentally verified subsequences which have been shown to contain a functional site in question. The regular expressions were manually derived form the alignments.

3.3.1.8 Evaluation of the regular expression

Regular expressions describing functional sites, are expected to overpredict as mentioned in section 1.2.1. But if the regular expression performs badly when exposed to protein sequences, countermeasures could be implemented. To monitor the performance of a regular expression, Scanprosite and elmfetch were used. Both approaches gave the same result: the number of hits a regular expression or pattern has in a protein database. Elmfetch was used for local searches with a regular expression in a protein database, and grep was used to count the hits. The online tool, ScanProsite, was also used for an online query. The regular expressions in this thesis were tested, and the results are shown in table 4.6. One countermeasure to improve the performance of a regular expression, was to divide one regular expression into several regular expres-

sions based upon homologous proteins. If, for example a regular expression was derived from 3 protein families, three separate regular expressions could be used instead. Often though, overprediction could not be reduced by altering the regular expression alone.

3.3.2 Collecting context information

3.3.2.1 Determination of taxonomic ranges

The protein(s) containing the recognition module was selected for the determination of the taxonomic range of the functional site[2]. To determine this, BLAST searches at ExPASy were performed against different taxonomic groups to see if the proteins containing the recognition module are present in a taxonomic group in question. For example, the Rb protein from human was used as a query against the following taxonomic groups: Metazoa (multicellular animals), Viridiplantae (green plants), *Saccharomyces cerevisiae* and *Plasmudium falciparum*. Only the Metazoa and Viridiplantae group showed significant hits, hence the functional range of Metazoa and Viridiplantae. The NCBI taxonomic identification numbers of Metazoa and Viridiplantae were retrieved using the NCBI taxonomic browser, and the corresponding taxonomy ID number was kept for future annotation. This method was used the in the determination of taxonomic ranges of all five functional sites in this thesis.

3.3.2.2 Determination of GO terms

The Gene Ontology (GO) terms are divided by the GO consortium into three different categories: molecular function, cellular component and biological process. One or more terms from each of these categories were manually selected to be associated with each functional site. No systematic method was used for the determination of GO terms. Manual navigation with the help of the GO browser, AmiGO, was used in navigating the GO tree, and appropriate GO terms were selected. The GO terms were selected based upon the collected knowledge of the recognition module, and not the proteins harboring the functional site. For example, the Rb proteins repress genes

[2]The functional site that interacts with the Rb proteins is also found in some viruses, which do not contain the Rb proteins. Making the deciding factor, distribution of recognition module, a bit questionable.

important in the S phase of the cell cycle. From this knowledge the following GO terms were selected: cellular compartment-*nucleus*; molecular function-*transcription regulator* and biological process-*regulation of cell growth*. All GO terms have a GO identification number and these numbers were stored for future annotation into the ELM database.

3.3.2.3 Discovering other protein units related to the functional site

A potential filter is the observed relationship a functional site has to another protein unit. To explore this possibility the SMART resource was used to predict domains in every protein sequence that have an experimental verified functional site. In this thesis no functional sites were found to be strictly associated with a specific domain, but in the case of the PxDLS functional site (see section 4.4), there is an over-representation of DNA-binding domains in the verified proteins. This observation can not be used as a filter, since some verified proteins do not have a DNA-binding domain, but can instead be used as information to strengthen predictions. Other relationships can be discovered from the biology of the functional site. This was the case of the RxL functional site (see section 4.3), where two functional sites (`ELM:LIG_CYCLIN` and `ELM:MOD_CDK`) are involved in the same biological process: cyclin/CDK dependent phosphorylation in the cell cycle [204]. This information could be used as a logical rule and hence a filter.

3.3.3 Annotation into the ELM database

A graphical interface was developed by members in the ELM consortium for the annotation of data into the ELM database. Following is the description of what kind of data was annotated. The following procedure was repeated for every functional site.

Before proceeding a following distinction should be mentioned. A functional site is the abstract *in vivo* functional unit. An ELM is a regular expression and contextual information tags representing a functional site. A functional site can be represented by several ELM's, but one ELM can only represent one functional site.

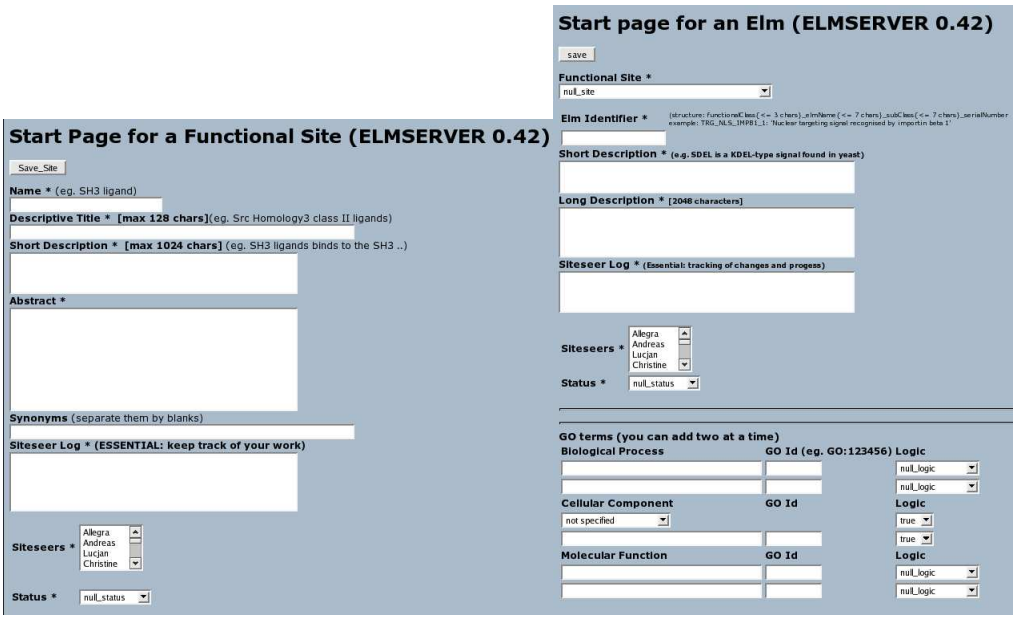


Figure 3.2: **Screen-shot from the ELM resource functional site and ELM input forms.** Functional site input form on the left. Here the name of the functional site, description and abstract were written. To the right is the ELM input form. Here a small description of the ELM, GO terms, NCBI taxonomy numbers and regular expression were annotated.

3.3.3.1 Functional site input

This is the first step in annotating a functional site into the ELM database. The name and an abstract were written and stored into the database for each of the five functional sites in this thesis, see Fig. 3.2.

3.3.3.2 ELM input

The ELM identifiers were defined for each functional site according to a ELM nomenclature defined by the ELM consortium. Beginning with the functional site class, followed by site name and then a serial number. For example the entry `ELM:LIG_RB`, shows that this functional site is a ligand and the name is here Rb. The name of the functional site are here used to reflect the proteins harboring the recognition module. If an additional functional site is showed to be a ligand to Rb proteins, this would have the nomenclature: LIG_RB_2. Determined NCBI identification numbers, from 3.3.2.1, and selected GO term identification numbers, from 3.3.2.2, were annotated for each functional site. The mature regular expression(s) was also annotated for each

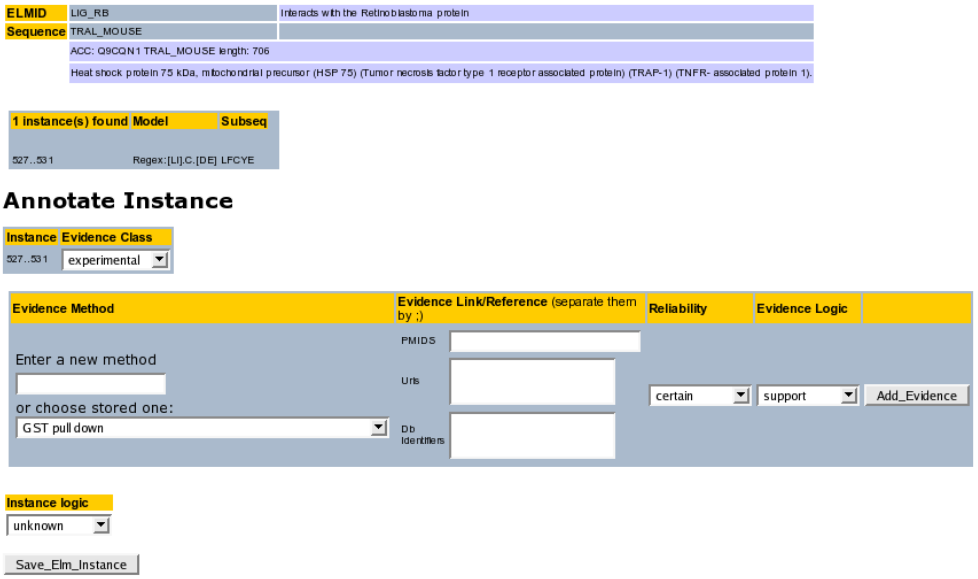


Figure 3.3: **Screen-shot from the ELM instance input interface.** First the acc. nr. / ID number of a protein were annotated. Then the input form searches for the selected pattern in the protein sequence and awaits conformation by the user. After conformation, appropriate experimental methods are annotated with the corresponding PubMed identification number.

functional site.

3.3.3.3 References input

Three references were selected for each functional site to be shown on the websites. These references were in many cases structural publications or good reviews of the functional site or the proteins harboring the recognition module. These references aid the users further into the scientific literature.

3.3.3.4 ELM instance input

This is the last step in the annotation of a functional site into the ELM database. This process begins with the annotation of a protein ID or accession number from section 3.3.1.3. The corresponding PubMed identification number, or PMID collected in 3.3.1.1 was then annotated. After this, determination of experimental method used in the PMID was determined and annotated. This was done for every single 97 proteins containing one of the five functional sites in this thesis. Following is a table which

shows the most frequent methods used in the publications[3]. The ID in table 3.3, are used in tables of verified proteins in Chapter 4. See Fig. 3.3. for a screen-shot of the ELM instance input interface.

Table 3.3: **Evidence codes**. Evidence codes used for annotation experimental evidence.

Method	Concept	ID
Pull down	Detection method	1
Yeast two-hybrid assay	Detection method	2
Co-immunoprecipitation	Detection method	3
Mutagenesis	Altering a protein	4
Structure	Detection method	5
Sequence similarity	Bioinformatics	6
Alanine scanning	Altering a protein	7
Western blot	Detection method	8
Motif deletion	Altering a protein	9
Sub-cellular localization	Detection method	10

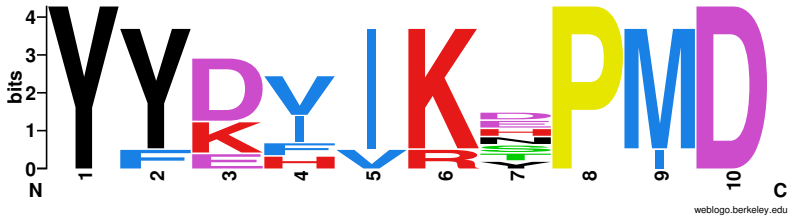
3.4 A brief introduction to regular expressions

Patterns. This is a brief overview of two similar ways to write patterns that are used to detect functional sites or other sequences features in protein sequences. A pattern is a way to describe rules that a sequence must fulfill in order to match the pattern, where the pattern is often related to a function. For a more extensive view [70].

Prosite patterns and regular expressions. Regular expressions are a language for matching textual patterns in strings, and is widely used in Unix and Unix like editors. The Prosite language is derived from regular expressions and has been developed an an independent syntax for biological applications [38]. A hypothetical example of going from a protein alignment to a regular expression representing rules which are considered important e.g. for an biological event, is shown in Fig 3.2. The pattern is a result of how a human or an algorithm *interprets* an alignment.

[3]This is not a comprehensive list of experimental methods in molecular biology. Instead the list is used to show common methods and experimental approaches in the collected publications in this thesis. Determining experimental approaches used in nearly 100 publications is difficult and annotations are prone to human error.

```
YYKIIKTPMD
YFKVIKHPMD
YYDIIKDPID
YYDFIKEPMD
YYEVIRSPMD
YYDHIKYPMD
YYDVVKNPMD
```




```
Y[FY]..[IV][KR].P[MI]D
```

```
Y-[FY]-x-x-[IV]-[KR]-x-P-[MI]-D
```

Figure 3.4: **Example of developing a regular expression.** Going first from an alignment, then identifying important residues and then describe the interpreted important residues in the form of a regular expression or a Prosite pattern. Middle picture made with Weblogo [190].

A real example: the functional site, di Lysine ER retrieving signal. The site is located at the C-terminus of ER type II membrane proteins and interacts with members of the coatomaer (COP I) [8]. It can be represented as a regular expression or a Prosite pattern, respectively; K.?K.{2,3}$ or K-x(0,1)-K-x(2,3)>. The sequences; KKEE, KGKALA and KKRGT match the patterns, whereas KKARGT does not. One can argue that the regular expression is easier to write and more difficult to read (for a human) and vice versa for the Prosite pattern, but the two are variants of a powerful tool for making rules and using these rules as a method for detecting subsequences with putative functions. Some aspects of regular expressions are discussed in section 5.1.

Another comprehensive system for describing consensus patterns has been developed, which aim to be a “normalization of symbols and terms used to describe, accurately and succinctly, the detailed interactions between residues of pairs of interacting proteins at protein:protein (or protein:peptide) interfaces” [1]. This system is currently not used in the ELM resource.

Chapter 4

Results

In the following chapter are the results of the siteseeing process of all five functional sites in this thesis. Among these, three functional sites are involved in the repression of transcription, where they act as mediators of protein-protein interactions when repressor complexes are recruited to genomic targets. These genomic targets include genes important in: embryonic development, cell growth, the cell cycle and cell homeostasis. One functional site which is involved in the formation of heterochromatin and long term transcriptional silencing of genes. The final functional site which is involved in the post-translational modification of proteins, regulating the activity of target proteins by phosphorylation. All have one thing in common: they are small subsequences found in a large number of proteins which are involved in important molecular processes in the cell.

The first five sections describe the five different functional sites analyzed in this thesis. Each section begins with an introduction to the biology of the functional site, then followed by a table of verified proteins containing the functional site. Then an alignment, Weblogo and regular expression(s) based on the collected proteins. Each section ends with comments about contextual information associated with each functional site. Section 4.6 shows how the regular expressions behave when exposed to protein sequences. Section 4.7 summarizes several new functional sites picked up during siteseeing and their current status. See also http://www.student.uib.no/~st04295/ for browsing the annotated functional sites in this thesis.

4.1 LxCxE

4.1.1 Introduction to Rb and the LxCxE functional site

The retinoblastoma protein (Rb) is a gene product of a tumor suppressor gene which is found to be inactivated in some cancers. The Rb protein and its relatives, p107 and p130, play a central role in controlling cell cycle progression from the G1 to the S phase of the cell cycle. The Rb proteins defines the restriction point, which is a time-point in the G1 phase where the cells are committed to enter the S phase. Prior to this point, the cell can take alternative routes such as differentiation, senescence, or cell death, depending on external signals received by regulatory proteins [154].

The fate of the cell in the cell cycle is determined by the phosphorylation state of the Rb proteins. This depends on cyclins/Cyclin Dependent Kinases (CDK) and CDK activating kinase (CAK). This control is due to the repression of the E2F transcription factors (E2F-1 through 5), which promotes expression of several genes that are required for the cell to enter the S phase. Some of these genes encode: cyclin A and E, dihydrofolate dehydrogenase, c-myb and c-myc [51]. The E2F transcription factors forms heterodimers with two transcription factors, DP-1 and DP-2, which are thought to stabilize E2F when bound to DNA [205]. In resting cells Rb actively represses E2F, but during the cell cycle the cell needs to express these E2F promoted gene products and Rb becomes hyper-phosphorylated by cyclin/CDK complexes. Phosphorylation by cyclin D/CDK4 and cyclin E/CDK2 of Rb proteins, releasing Rb and exposing E2F. Due to their critical role in repression and control of the cell cycle, the Rb proteins are targeted by several viruses which exploit the cell replication machinery by pushing the cell into S phase, to promote viral growth and replication. This is achieved by expressing oncoproteins that bind and inactivates Rb proteins and thus inactivate repression similar to phosphorylation-mediated inactivation of Rb [205].

The Rb proteins have two protein-protein interaction “pockets”, which allows Rb to be bound to E2F and a repressor (or a viral oncoprotein) simultaneously. The E2F binding pocket is denominated the “A/B pocket”, since the interaction interface is defined by two cyclin-like domains: the A and B domains. The other pocket is named also after the domain where its located, the “B pocket”. This B-pocket has evolved into

a location for protein-protein interaction to take place, where a common motif found in every protein which interacts with the B-pocket of Rb proteins.

The protein-protein interaction motif found in Rb-interacting is denominated, “the LxCxE motif”, after the pattern initially thought to be required for the interaction. Rb mediated repression requires that both the A/B pocket and the B-pocket are intact and functional so that Rb can dock repressors, via the B-pocket, and target repression at the E2F promoter, via the A/B pocket. The major role of the LxCxE binding site in Rb is to be a docking site for proteins (co-repressors or bridging proteins) that subsequently recruit repressors like: histone deacetylase proteins, histone methyltransferases and chromatin remodeling complexes. This results in Rb and bridging proteins like CtIP [148] the RBP1 [117] proteins co-operate to form a repressional complexes.

4.1.2 LxCxE containing proteins

Table 4.1[1] on page 42, summarizes all collected experimentally verified instances that uses the LxCxE motif in interaction with Rb proteins. From looking at the description column in table 4.1, it is apparent that there are many different types of proteins that contain the LxCxE functional site, ranging from a ubiquitin hydrolase to a RNA-specific adenosine deaminase. Presumably most proteins are co-repressors, like CtIP and RBP1, which in turn recruits repressors. Fig 4.2 shows the alignment of the LxCxE functional site and 10 additional residues on each side. Notice the diversity of the amino acids flanking the functional site. One protein has it in the C-terminal and another protein has it in the N-terminal. Some proteins have several proline and glycine residues (which are expected since functional sites are thought to reside in non-globular regions), and in contrast some are lacking flanking prolines and glycines. There are no obvious similarity besides the motif itself.

[1]The table layout is repeated for all five functional sites in this thesis. ID/Acc. is the ID or accession number from annotated proteins in databases in UniProt. The short descriptions are directly taken from UniProt annotations. * is used to show that this is a Pfam domain clash. The numbers in the “Method” column corresponds to evidence codes used in table 3.3.

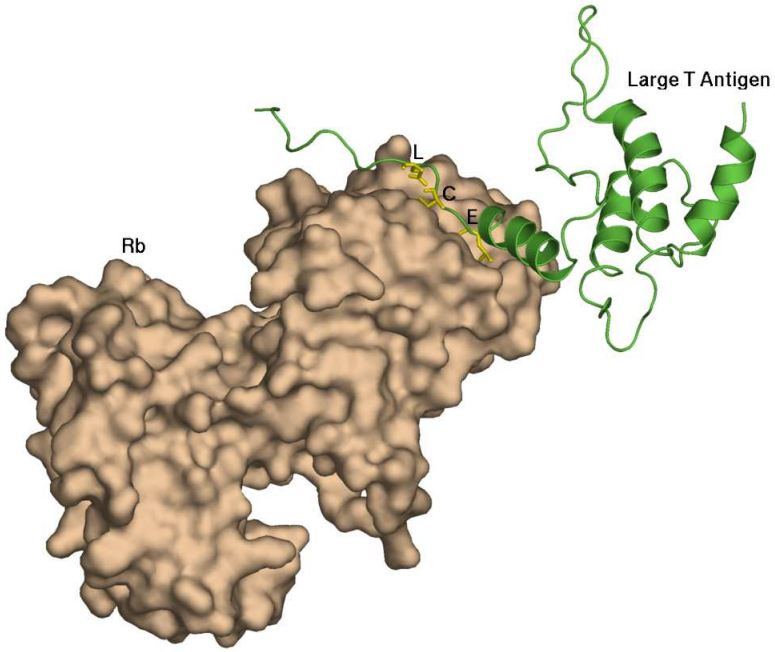
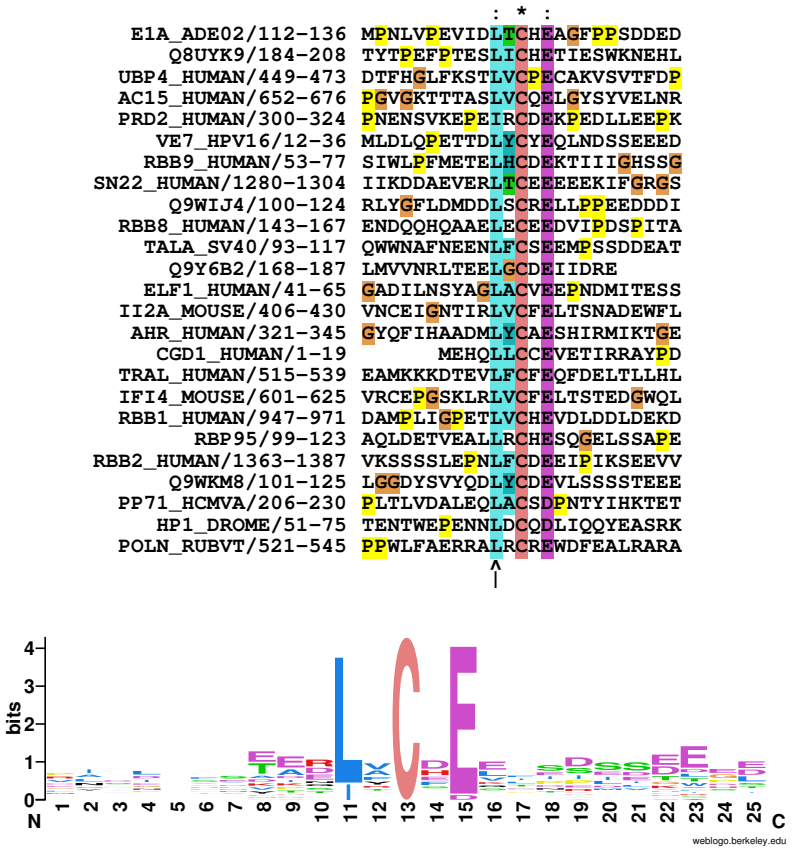


Figure 4.1: **Structure of LxCxE motif bound to its recognition module in Rb.** The LxCxE motif in the Simian Virus 40 large T antigen protein (green cartoons) is shown bound to the retinoblastoma protein (shown in wheat color and surface representation). In retinoblastoma, the A domain is the globular domain on the left and the B domain is the globular domain on the right. The B domain is the recognition module for the LxCxE motif. PDBid: 1GH6, [109].



[LI].C.[DE]

Figure 4.2: **Alignment of the motif from proteins containing the LxCxE motif.** Alignment, Weblogo and regular expression of verified proteins in table 4.1.

4.1.3 Context rules and filters

Localization. Most proteins that interact with Rb are associated and localized in the nucleus, but at least one protein, hsp75, is localized in the cytoplasm. The hsp75 protein has been proposed to act as a molecular chaperone for the Rb proteins, aiding in refolding to their native state, after phosphorylation by cyclin/CDK [41]. Although the LxCxE binding function is regulated by phosphorylation at T821 and T826 in Rb [112], the interaction may depend on de-phosphorylation by PP1-alpha, to allow interaction between Rb and hsp75 [184]. The interaction between these two must therefore be limited to specific time points in the cell cycle (breakdown of the nuclear envelope) or another special condition where the two proteins moves from their usual compartment.

Although Rb has been shown to interact with a non-nuclear protein, hsp75, the annotation of the LxCxE motif to the nucleus is justified by the majority of verified proteins are localized in the nucleus, and the application of the nucleus as a filter (Table 5.1.).

Taxonomy. The retinoblastoma proteins Rb, p130 and p107 were first found in vertebrates, but homologs has been identified in other metazoans and plants, like *C. elegans* [137], *Z. mays* (maize) [75], and *A. thaliana* [106], and in the green algae, *Chlamydomonas* [212]. Plant and metazoan Rb proteins have been shown to from repressor complexes and are susceptible to several viral oncoproteins that utilize the LxCxE motif, e.g. bean yellow dwarf virus and banana bunchy top virus [221], suggesting a similar function *in vivo*. Although Rb has been identified in the algae *Chlamydomonas*, its is unclear if the protein share function with its metazoan relatives. Retinoblastoma binding proteins that uses the LxCxE functional site are at least found in: *metazoa* (multicellular animals, Taxonomy ID: 33208), *viridiplantae* (green plants, Taxonomy ID: 33090) and viruses (Taxonomy ID: 10239).

Globular clash filter. 5 of the 17 verified cases (table 4.1) have the functional site inside a domain: RFC1_HUMAN (`SMART:AAA`)[2], AHR_HUMAN (`SMART:PAS`), II2A_MOUSE

[2]This refers to a SMART, Pfam or ELM entry .

and IFI4_MOUSE (`Pfam:HIN`), TRAL_HUMAN (`Pfam:HSP90`) (results not shown).

Although the globular filter removes five true instances for the LxCxE motif, it shows itself to be a powerful filter [180].

Molecular context and co-occurrences. The domain architecture of verified proteins that uses the LxCxE functional site are diverse. There are no other co-occurrences of protein units that can be associated with the LxCxE functional site.

Table 4.1: **Verified proteins which uses the LxCxE functional site**.

Name	ID/Acc.nr.	Short description	Length	Dom.	PMID	Ref.	Method
RBP1	RBB1_HUMAN	Retinoblastoma-binding protein 1	1257	-	8414517	[117]	3
RBP2	RBB2_HUMAN	Retinoblastoma-binding protein 2	1722	-	8414517	[50]	3
RFC1	RFC1_HUMAN	Replication factor C large subunit	1147	yes	11336696	[173]	1
IFI2	Il2A_MOUSE	Interleukin-2 receptor alpha chain precursor	445	yes*	7890747	[44]	4
IFI4	IFI4_HUMAN	Double-stranded RNA-specific adenosine deaminase	640	yes*	10951565	[88]	2
UBP	UBP4_HUMAN	Ubiquitin carboxyl-terminal hydrolase 4	963	-	11571651	[31]	3
ELF-1	ELF1_HUMAN	ETS-related transcription factor Elf-1	619	-	8493578	[219]	1
AHR	AHR_HUMAN	Aryl hydrocarbon receptor precursor	848	yes	10644764	[59]	2
BRM	SN22_HUMAN	Possible global transcription activator SNF2L2	1586	-	9326598	[209]	3
RBP95	Q9HC82	Rb-associated protein	838	-	10944455	[223]	2
HSP75	TRAL_HUMAN	Heat shock protein 75 kDa	704	yes	8756626	[41]	4
BOG	RBB9_HUMAN	Retinoblastoma-binding protein 9	186	-	9697699	[227]	2
EID	Q9Y6B2	RB-and P300-binding protein EID-1	187	-	11073990	[142]	4
Cyclin D1	CGD1_HUMAN	G1/S-specific cyclin D1	295	-	7696881	[32]	2
RIZ	PRD2_HUMAN	PR-domain zinc finger protein 2	1719	-	7538672	[39]	4
CtIP	RBB8_HUMAN	Retinoblastoma-binding protein 8	897	-	10449734	[148]	2
HP1	HP1_DROME	Heterochromatin protein 1	206	-	11533237	[216]	1
Viral							
HPV E7	VE7_HPV16	E7 protein	98	yes*	9495340	[123]	5
BBTVDNA5	Q9WKM8	Hypothetical protein	129	-	10640570	[221]	2
NSP90	POLN_RUBVT	RNA-directed RNA polymerase/helicase	2205	yes*	10073691	[68]	4
SV40 largeT	TALA_SV40	Large T antigen	708	yes*	11226179	[110]	5
BYDV RepA	Q8UYK9	Replication-associated protein A	264	-	10191192	[132]	2
Clink	Q9WIJ4	Cell cycle link (CLINK) protein	169	-	10708410	[13]	4
A E1A	E1A_ADE02	Early E1A 32 kDa protein	289	yes*	2538790	[40]	1
PP71	PP71_HCMVA	71 kDa upper matrix phosphoprotein	559	-	12612064	[104]	4

4.2 SID

4.2.1 Introduction to Sin3 and the SID functional site

The SID (Sin3 Interacting Domain) was first found in the Mad family of transcriptional repressors [19]. The Mad proteins, are members of the Myc/ Max/ Mad network which regulates different genes involved in several aspects of cell behavior (reviewed in [139]). This SID interacts with the PAH2 (Paired Amphipatic repeat 2) domain, which is one of three (PAH1, PAH2 and PAH3) protein interaction domains in the paralogs Sin3A and Sin3B in mammals. The Sin3 proteins are thought to be scaffold proteins. They have multiple interaction partners and form a transcriptional repressor complex called the Sin3-complex, which has a central role in Sin3-dependent transcriptional repression in diverse cellular processes such as proliferation, differentiation, apoptosis and cell fate determination. Components of this complex include two histone deacetylase proteins (HDAC1 and HDAC2), RbAp46, RbAp48, SAP18, SAP30 and the Sin3 protein itself (reviewed in [5]). In addition, several new components of the Sin3 complex has recently been identified: SAP45, SAP130 and SAP180 [67].

The SID itself is an amphipatic α- helix (hydrophobic on one side and hydrophilic on the other, in respect to the axis) and interact with a deep hydrophobic groove defined by the PAH2 domain in Sin3B shown in Fig 4.3. The SID is found in two different protein families which repress transcription: 4 members of the Mad family and 5 members of the sp-1 like/KLF transcription factor family. In addition, a yeast protein (UME6) and a PHD domain containing protein (Pf1) has show to contain the a SID .

4.2.2 SID containing proteins

The proteins that have experimental evidence for the use of the SID helix in interaction with the PAH2 domain of Sin3 are listed in table 4.2. All verified proteins are transcription regulators which recruits the Sin3 repressor complex via the SID functional site. The alignment of all SID containing proteins shows, to some extent, that the SID shares little sequence similarity between the different proteins. Orthologs are included due to a low number of verified proteins. The pattern was divided up into three differ-

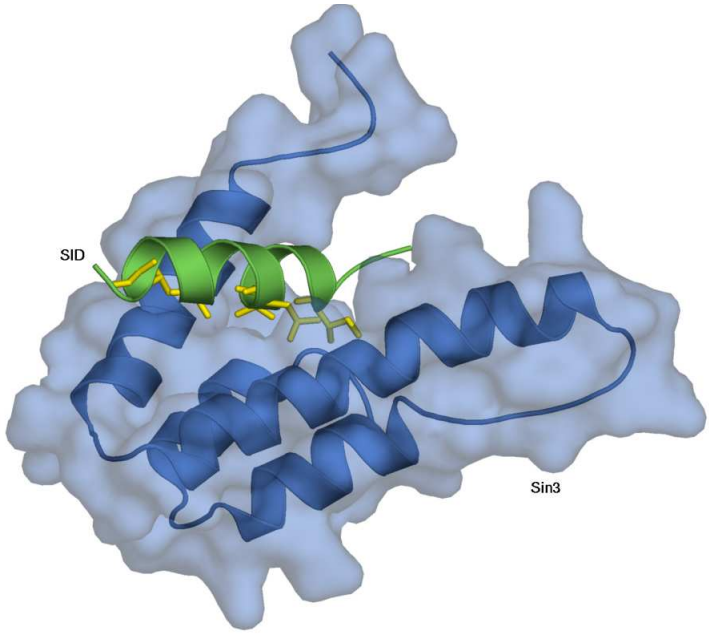


Figure 4.3: **Structure of the SID α-helix bound to the PAH2 domain of Sin3B.** The PAH2 domain of Sin3B (colored blue and shown in cartoon and transparent surface) and the SID helix from Mad1 (green cartoon). The important amino acid residues in the SID helix are shown in yellow sticks. PDBid: 1G1E, [37].

ent regular expressions, due to a to weak performance of a single regular expression. A proposal to one single regular expression is: `[LMYMHFP]..[AVF]..[FYLA][LMVA]`. This expression has more then one hit pr. protein in Swiss-Prot. To prevent this, the single regular expression was divided into three different regular expressions. One expression reflecting the Mad proteins, one expression representing the sp-1 like proteins and one regular expression reflecting the unrelated proteins, TGIF, UME6 and Pf1. This difficulty in making one single regular expression may be the result of a poor alignment.

4.2.3 Context rules and filters

Localization. All SID containing proteins are transcription factors that are localized in the nucleus.

Taxonomy. Proteins that uses the SID are found in *H. sapiens*, *M. musculus* and *S. cerevisiae*. To investigate if the Sin3 protein is present in plants a BLAST search was

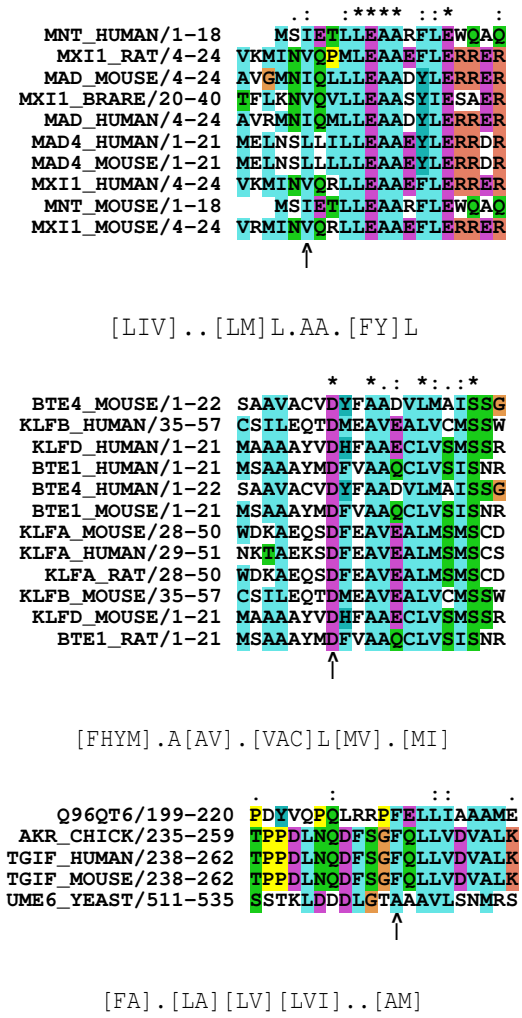


Figure 4.4: **Three alignments of the SID helix from proteins listed in table 4.2.** The dataset was divided into three groups, depending on either they belong to the mad family, the sp-1 like family or neither. Top: MAD family. Middle: sp-1 like proteins. Bottom: The unrelated proteins Pf1, TGIF and UME6. Weblogo is not shown.

performed as described in 3.3.2.1. Several viridiplantae proteins showed significant hits. Hence presuming taxonomic range of *Eukaryota*; Taxonomy ID: 2759

Globular clash filter. The SID has not been observed inside a domain sequence.

Co-occurrence and molecular context. Most SID containing proteins have a domain that has a function related to transcription. The Mad proteins have a Helix-loop-Helix dimerization domain (SMART:HLH), sp-1 like has 3 DNA binding Zinc fingers (SMART:ZnF_C2H2) the Tgif has a DNA binding homeo domain (SMART:HOX), and the UME6 yeast protein has a DNA binding domain of the Gal4 family (SMART:Gal4). The Pf1 protein has two PHD fingers (SMART:PHD) which are associated to several functions including chromatin binding activity [181] and as a phosphoinositide receptor [74].

Table 4.2: **Verified proteins which uses the SID functional site.**

Name	ID/Acc.nr.	Short description	Length	Dom.	PMID	Ref.	Method
MAD	MAD_HUMAN	MAX dimerizer	221	-	11101889	[201]	5
MXI1	MXI1_HUMAN	MAX interacting protein	228	-	10918583	[150]	8
MNT	MNT_HUMAN	MAX binding protein MNT	582	-	9000049	[93]	2
MAD4	MAD4_HUMAN	Max-interacting transcriptional repressor MAD4	209	-	8816491	[19]	6
TIEG1	KLFA_HUMAN	Krueppel-like factor 10	480	-	11438660	[233]	1
TIEG2	KLFB_HUMAN	Krueppel-like factor 11	512	-	12006497	[60]	1
BTEB3	KLFD_HUMAN	Krueppel-like factor 13	288	-	11477107	[102]	1
BTEB1	BTE1_HUMAN	Krueppel-like factor 9	244	-	11438660	[233]	1
BTEB4	BTE4_HUMAN	Krueppel-like factor 16	252	-	11438660	[233]	1
TGIF	TGIF_HUMAN	5’-TG-3’ interacting factor	272	-	11571228	[228]	8
UME6	UME6_YEAST	Transcriptional regulator UME6	836	-	9150136	[103]	4
Pf1	Q96QT6	PHD zinc finger transcription factor	704	-	11390640	[230]	2

4.3 RxL

4.3.1 Introduction to cyclin and the RxL functional site

Cyclins are a family of proteins that regulate the activity of a group of protein kinases: the cyclin-dependent kinases (CDK) by binding and forming cyclin/CDK hetrodimers. In addition to cyclin binding, a specific threonine residue (T161) in CDK must be phosphorylated by CAK (CDK activating kinase), representing the final step in CDK activation [64]. The cyclin/CDK complexes phosphorylate different proteins in order to orchestrate the complex task of regulating the different stages in the cell cycle. Each cyclin have its specific CDK binding partner(s), and both cyclin and CDKs are expressed and degraded at different stages during the cell cycle. In humans, 8 different CDKs and 10 different cyclins have been identified and some of the observed combinations are: cyclin D/CDK4/6 (G1 phase), cyclin E/CDK2 (G1/S transition), cyclin A/CDK2 (S and G2 phase) and cyclin B / CDK1 (G2 and M phase) [153].

There are two functional sites associated with cyclin/CDK dependent phosphorylation: a CDK consensus phosphorylation site and a cyclin interaction site. This last site, the Cy motif or RxL motif, has been show to interact with cyclins and thereby increase the level of phosphorylation of a CDK substrate. This requires that a substrate of a cyclin/CDK complex are in proximity of both proteins, and these two sites should be separated by at least 12 amino acids [204]. Insertion of a RxL motif in surface loops has been observed to increase phosphorylation of otherwise non-phosphorylated substrates [55]. The structure of the cyclin A/CDK2 with several peptides containing the RxL motif from proteins like p53, Rb, E2F, p107 and the inhibitor, p27, have been solved [135]. In addition, a motif found in Xenopus cyclin B2, the RRASK motif, has been implicated in the substrate recognition of cyclins [73], and may be a variant of the RxL motif.

4.3.2 RxL containing proteins

The RxL functional site has been identified in several eukaryotic proteins (table 4.3). The majority of proteins in table 4.3 are transcription factors and co-repressors, like E2F, p53 and Rb. In addition, several inhibitors of cyclin/CDK activity uses the motif

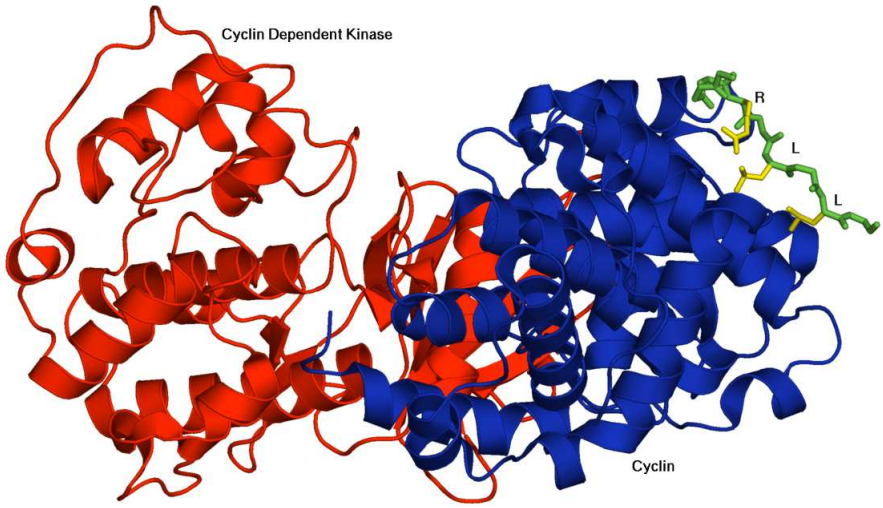


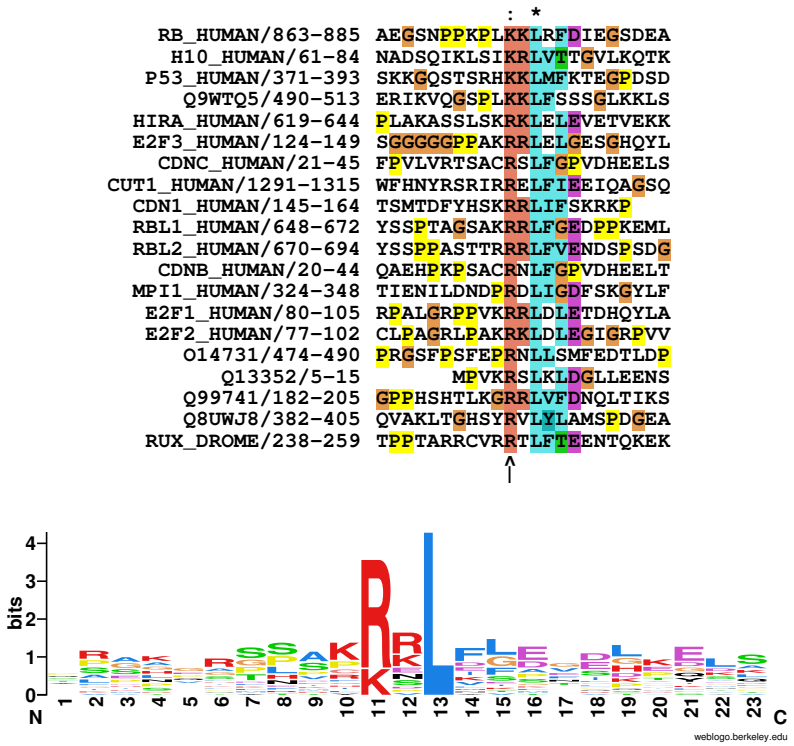
Figure 4.5: **Structure of a RxL containing peptide bound to a cyclin/CDK hetrodimer.** Cyclin A (blue cartoon)/ CDK2 (red cartoon) hetrodimer bound to a RxL containing peptide derived from E2F (backbone shown in green sticks). The amino acid residues defining the functional site are shown in yellow sticks. PDBid: 1H24, [135].

in interaction with cyclin, like p21, p27 and roughex. The alignment of the motif is shown in figure 4.6. At first glance only the consensus (RK)xL are common to all sequences. But further analysis shows hydrophobic residues are common in the first or second residue after the (RK)xL consensus sequence. Structural data obtained by [135] shows this hydrophobic residue does indeed contribute to significant van der Waals interaction between the peptide and cyclin A. These have therefore been included in the regular expression, which then becomes `[RL].L.{0,1}[FYLIVMP]`, where `.{0,1}` means that any are allowed to occupy this position, but none is required. Also notice the frequency of proline and glycine residues flanking the functional site.

Different cyclins/CDKs presumably have different substrates. It is thus possible that the RxL motif is the determinant for specificity. Future work will hopefully determine if the RxL functional site should be divided into several functional sites, each with their specific cyclin binding partner.

4.3.3 Context rules and filters

Localization. The cellular localization of cyclin/CDK complexes are diverse. Some cyclin/CDK complexes are primarily localized to the nucleus, other to the cytoplasm



[RK].L.{0,1}[FYLIVMP]

Figure 4.6: **Alignment of the verified proteins carrying the RxL functional site listed in table 4.3.** Alignment, Weblogo and regular expression of all experimentally verified proteins which uses the RxL motif.

and yet other shuttle in and out of the nucleus. Localization of cyclin/CDK complexes are reviewed in [229]. The major cytoplasmic cyclin/CDK proteins are the cyclin B1 and B2/CDK1 (formerly called Cdc2) complexes. In addition, cyclin D1/CDK2/4 are exported from the nucleus to the cytoplasm during the G1 to S transition. The activity of cyclin/CDK complexes and the RxL motif is thus limited to the nucleus and cytoplasm.

Taxonomy. Cyclins are found in all eukaryotic organisms, and presumably so is the RxL motif, hence the annotation *Eukaryota*: Taxonomy ID: 2759. The RxL functional site has, however, not yet been identified in lower eukaryotes.

Globular clash filter. The non-globular filter rules out two known proteins, the (CDP)/Cux transcription factor and the Cdh1 protein. The site in (CDP)/Cux are inside a homeo domain (`SMART:HOX`) and is located at the very end of the domain, still suggesting accessibility. In the other protein, Cdh1, is found to be in the middle of a WD40 repeat domain (`SMART:WD40`) (results not shown).

Molecular context and co-occurrences. As mentioned earlier, a substrate for phosphorylation of a cyclin/CDK complex presumably should have two functional sites: the RxL motif and the CDK phosphorylation consensus motif, (ST)Px(KR). Since both are annotated in the ELM resource, there is an opportunity that they should be associated with one another and thereby provide a co-occurrence filter.

Table 4.3: **Verified proteins which uses the RxL functional site.**

Name	ID/Acc.	Short description	Length	Dom.	PMID	Ref.	Method
Rb	RB_HUMAN	Retinoblastoma-associated protein	928	-	9891042	[3]	9
p107	RBL1_HUMAN	Retinoblastoma-like protein 1	1068	-	11884610	[124]	4
p130	RBL2_HUMAN	Retinoblastoma-like protein 2	1139	-	11884610	[84]	6
E2F1	E2F1_HUMAN	Transcription factor E2F1	437	-	9199321	[55]	1
E2F2	E2F2_HUMAN	Transcription factor E2F2	437	-	9199321	[55]	6
E2F3	E2F3_HUMAN	Transcription factor E2F3	465	-	9199321	[55]	6
HIRA	HIRA_HUMAN	TUP1 like enhancer of split protein 1	1078	-	11238922	[81]	3
Cux	CUT1_HUMAN	CCAAT displacement protein	1505	-	11584018	[187]	1
Myt1	O14731	Membrane-associated kinase	499	-	10373560	[131]	9
p53	P53_HUMAN	Cellular tumor antigen p53	393	yes*	10884347	[138]	3
SSeckS	Q9WTQ5	PKC binding protein SSECKS	1684	-	10982843	[127]	4
Roughex	RUX_DROME	Cell cycle negative regulator roughex	335	-	11027291	[18]	2
b3-endonexin	Q13352	Beta 3-endonexin	170	-	10673397	[166]	1
CDC6	Q99741	CDC6-related protein	560	-	9889196	[175]	1
cdc25a	MPI1_HUMAN	M-phase inducer phosphatase 1	523	-	9234691	[186]	4
p202	Q13632	NPAT protein	1175	-	10995387	[141]	10
p21	CDN1_HUMAN	Cyclin-dependent kinase inhibitor 1	164	yes*	11438644	[42]	4
p27	CDNB_HUMAN	Cyclin-dependent kinase inhibitor 1B	198	yes*	9488039	[4]	4
p57	CDNC_HUMAN	Cyclin-dependent kinase inhibitor 1C	316	yes*	10713702	[85]	9
Cdh1	Q9UI96	Fizzy-related protein homolog	453	yes	11340163	[199]	8

4.4 PxDLS

4.4.1 Introduction to CtBP and the PxDLS functional site

The CtBP, or C-terminal Binding Protein, was first discovered as an interaction partner of the viral protein E1A, expressed by all human adenoviruses. Further analysis showed that only a small subsequence in the viral protein was sufficient for interaction with CtBP [188]. After this discovery several eukaryotic proteins involved in transcriptional repression has been identified to interact with CtBP proteins via this small motif including the mouse krüppel-like factor (BKLF) [210]. In vertebrates two paralogs of CtBP have been identified: CtBP1 and CtBP2. These are able to form homo- and heterodimers and thereby may be able to bind two motif containing proteins simultaneously [193]. The CtBP binding proteins which uses the motif can be divided into two groups: one group are proteins associated with DNA (direct or indirect DNA binding), or the other group of proteins associated with transcriptional repression (histone deacetylase proteins type II, polycomb 2). This suggests that the CtBP proteins acts as a bridge, recruiting repressors to DNA bound proteins.

The CtBP proteins share a high degree of similarity to a group of dehydrogenases called, 2-hydroxy dehydrogenases. The interaction between CtBP and a motif-containing protein has been shown to depend on NAD+/NADH binding [115].

4.4.2 PxDLS containing proteins

A large number of PxDLS containing proteins have been identified and are listed in Table 4.4. The verified proteins can be roughly be divided into two groups: either DNA bound transcription factors or repressors like the histone deacetylase proteins (which may not be surprising since the role of CtBP is to recruit repressors to transcription regulators). Following is the alignment of the subsequences of the proteins in table 4.4. Firstly the abundance of proline and glycine is obvious and suggests that these functional sites reside in non-globular regions. The first residue of the motif is a proline or glycine, followed by mostly hydrophobic residues and two cases where glutamate is present. In the middle of the functional site is there preference for the acidic amino acids glutamate and aspertate but two instances each of aspergine and serine

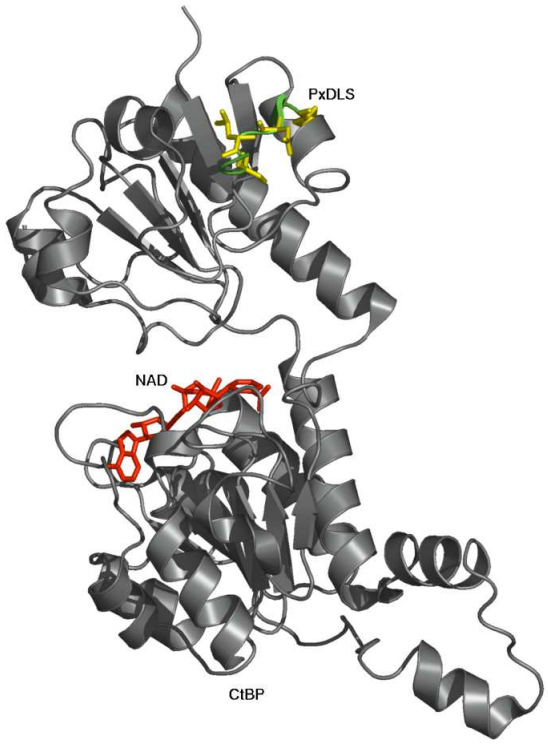
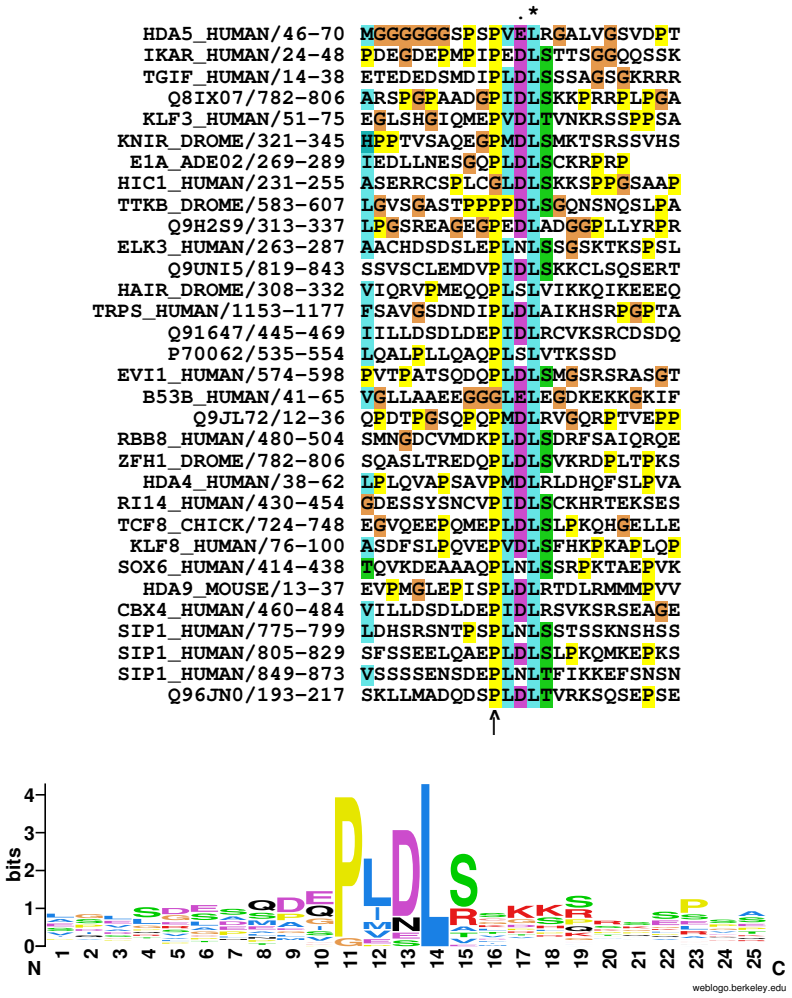


Figure 4.7: **Structure of the PxDLS motif bound to the CtBP protein from rat.** The CtBP/BARS protein (gray cartoon) bound to a PxDLS containing peptide (green sticks) and NADH (red sticks). The important residues in the PxDLS containing peptide are displayed in yellow sticks. PDBid: 1HL3, [162].

occurs, making this position hydrophilic. In the next position the leucine is completely conserved. In the last position, there is predominance of hydrophilic amino acids, but two instances of valine is observed. It has been suggested that a lysine residue, followed 1-4 amino acids C-terminal to the motif, is involved in regulation of the PxDLS functional site. Mutational analysis of the viral protein E1A, that mimic acetylation, where lysine is mutated to glutamine, resulted in proteins that were defective in CtBP binding [234].

4.4.3 Context rules and filters

Localization. CtBP is localized in the nucleus and is associated with transcriptional repression. Remarkably, CtBP has also been shown to be involved in tubule fission in the Golgi. This dual function is proposed to be dependent on the ligand bound to the CtBP proteins, where CtBP bound to NAD+/NADH is associated with repression in the



[PG][LVIPME][DENS]L[VASTRGE]

Figure 4.8: **Alignment of proteins listed in table 4.4.** Alignment, Weblogo and regular expression of proteins listed in table 4.4. of all experimentally verified proteins which uses the PxDLS motif in interaction with the CtBP protein.

nucleus and CtBP bound to acyl-CoA has a role in the Golgi [222, 162]. Presumably a CtBP protein bound to acyl-CoA is unable to bind to a PxDLS containing protein, since PxDLS binding has been shown to be NAD+/NADH dependent. Hence all eukaryotic proteins that uses the PxDLS motif in interaction with the CtBP proteins seems to be associated with repression in the nucleus.

Taxonomy. The CtBP proteins have been found in both *Mus musculus* (mouse), *Homo sapiens* (human), *Drosophila melanogaster* (fruitfly), *Xenopus laevis* (frog), *Arabidopsis thaliana* (mouse-ear cress) and *Danio rerio* (zebrafish). Although *Arabidopsis thaliana* does have a CtBP homologue, it is unclear if the protein has PxDLS binding capability. PxDLS binding activity has been shown for CtBP proteins in *Drosophila melanogaster* [177], *Xenopus laevis* [193], and *Homo sapiens* [188]. The taxonomy range for the PxDLS functional site is therefor *Metazoa;* Taxonomy ID: 33208.

Globular clash filter. There is only one instance where the PxDLS functional site is found inside a SMART domain. This is in the actin domain (`SMART:ACTIN`) of mArpα, but the primary sequence and the occurrence of glycine flanking the functional site suggests that the functional site resides in a loop of the domain (results not shown).

Molecular context and co-occurrences. Strikingly 18 of the 30 experimentally verified proteins contain domains that are associated with DNA: 14 proteins contain multiple Zn-fingers, one protein has a ETS domain (Net), one protein has a homeo domain (TGIF) and two proteins have a chomo domain (the polycomb homologous; xpolycomb and HPc2) . Four of the instances are histone deacetylase type II proteins, making 22 of 30 verified instances either with DNA associated domains or proteins with histone deacetylase activity. The remaining 8 proteins may be adapter proteins (like CtIP) linking histone deacetylase proteins to other DNA bound proteins.

Table 4.4: **Verified proteins which uses the PxDLS functional site.**

Name	ID/Acc.nr	Short description	Length	Dom.	PMID	Ref.	Method
CtIP	RBB8_HUMAN	Retinoblastoma-binding protein 8	898	-	9535825	[189]	3
BKLF	KLF3_HUMAN	Kruppel-like factor 3	345	-	10756197	[210]	6
KLF8	KLF8_HUMAN	Krueppel-like factor 8	359	-	10756197	[215]	1
Fog-1	Q8IX07	Zinc finger protein ZFPM1	1004	-	10995736	[107]	4
Fog-2	Q9UNI5	Zinc finger protein ZFPM2	1151	-	11940669	[107]	6
AML1	EVI1_HUMAN	Ecotropic virus integration 1 site protein	1051	-	11965542	[97]	9
Ikaros	IKAR_HUMAN	DNA-binding protein Ikaros	519	-	10766745	[113]	8
ZFH-1	ZFH1_DROME	Zinc finger homeodomain protein 1	1060	-	10359772	[178]	8
Ef1 delta	TCF8_HUMAN	Transcription factor 8	1124	-	10567582	[71]	2
Tgif	TGIF_HUMAN	5'-TG-3' interacting factor	272	-	10995736	[147]	7
Hairy	HAIR_DROME	Hairy protein	337	-	9524128	[177]	2
Net	ELK3_HUMAN	ETS-domain protein Elk-3	407	-	10369679	[48]	2
Knirps	KNIR_DROME	Zygotic gap protein knirps	429	-	9843507	[108]	4
Rip 140	RI14_HUMAN	Nuclear receptor interacting protein 1	1158	-	11509661	[218]	2
HDAC4	HDA4_HUMAN	Histone deacetylase 4	1084	-	11022042	[232]	6
HDAC5	HDA5_HUMAN	Histone deacetylase 5	1122	-	11022042	[232]	6
HDAC7	Q9JL72	Histone deacetylase 7a	938	-	11022042	[232]	6
MITR	HDA9_MOUSE	Histone deacetylase 9	588	-	11022042	[232]	3
xpolycomb	Q91647	XPolycomb	521	-	9858600	[193]	2
HPc2	CBX4_HUMAN	Chromobox protein homolog 4	558	-	9858600	[193]	2
Eos	Q9H2S9	Zinc finger transcription factor Eos	483	-	12444977	[174]	2
Trps1	TRPS_HUMAN	Zinc finger transcription factor Trps1	1281	-	12449777	[174]	2
ArpN alpha	B53B_HUMAN	53 kDa BRG1-associated factor B	426	yes	12565893	[168]	2
LcoR	Q96JN0	Hypothetical protein	572	-	12535528	[63]	4
Tramtrack	TTKB_DROME	Tramtrack protein, beta isoform	643	-	10978285	[224]	7
XTcf-3	P70062	Transcription factor XTCF-3	554	-	10375506	[34]	1
SOX6	SOX6_HUMAN	Transcription factor SOX-6	828	-	11504872	[157]	2
HIC1	HIC1_HUMAN	Hypermethylated in cancer 1 protein	733	-	12052894	[52]	2
SIP1	SIP1_HUMAN	Zinc finger homeobox protein 1b	1214	-	12714599	[213]	4
Viral Ade. E1A	E1A_ADE02	Early E1A 32 kDa protein	289	-	7479821	[188]	4

4.5 PxVxL

4.5.1 Introduction to HP1 and the PxVxL functional site

The HP1 proteins (heterochromatin protein 1) are non-histone chromosomal proteins associated with heterochromatin. Heterochromatin is a highly condensed state of chromatin associated with silencing of genomic regions. Genes which resides in a heterochromatic region are not transcribed and are replicating late. There are two types of heterochromatin, constitutive and facultative. Constitutive heterochromatin is found at distinct chromosome territories such as pericentric and telomeric regions. Facultative heterochromatin is more dynamic and regulated regions where genes can be switched on and off, depending on e.g. development stages of an organism. In both types of heterochromatin, the HP1 proteins are central in its formation [46]. The role of HP1 proteins in silencing is thought to be widespread in the genome and the mechanism is conserved from yeast to man [220].

In mammals, there are three members of the HP1 family: HP1α, HP1β and HP1γ, consisting of an N-terminal chromo domain [194], a variable hinge region and a C-terminal chromo shadow domain [2]. The chromo domain in HP1 proteins has been reported to interact with the N-terminal histone tail in H3 at a specific methylated lysine residue, K9 [99]. The chromo shadow domain have been shown to bind to a consensus peptide pentamer, and thereby suggesting that the chromo shadow domain mediate protein-protein interactions [196]. Further NMR analysis suggested that the chromo shadow domain forms a homodimer, and a consensus peptide interacts with an interface defined by the chromo shadow dimer [35]. This was confirmed by a final NMR analysis of the chromo shadow homodimer bound to a PxVxL peptide [206], shown in Fig 4.9.

4.5.2 PxVxL containing proteins

Table 4.5. shows all experimentally verified proteins shown to use the PxVxL functional site in interaction with HP1. Common to the proteins is that they are all associated with DNA, either this is to add the methyl group to the histone tails of the nucleosomes, like the histone methyltransferase Su(var)3-9 or to initiate transcription

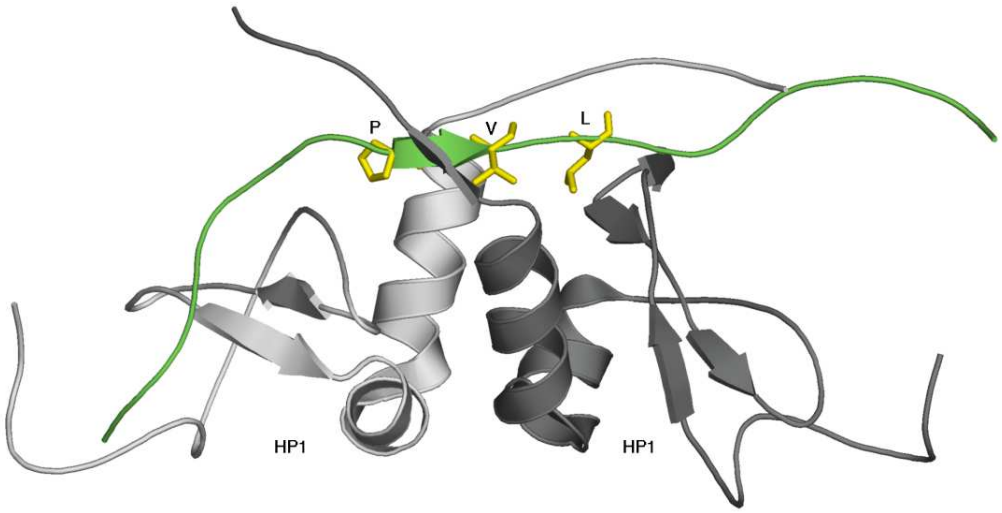
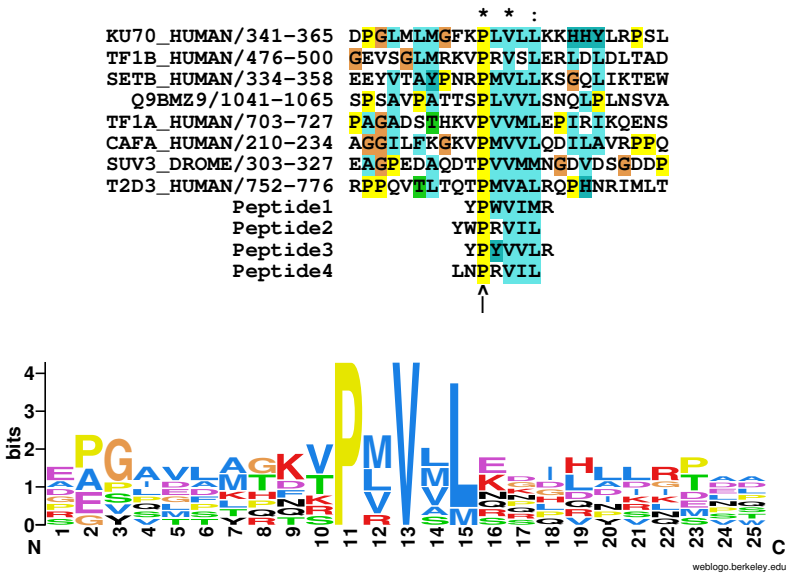


Figure 4.9: **Structure of a PxVxL containing peptide derived from the Chromatin Assembly Factor 1 protein (CAF-1) bound to a chromo shadow homodimer.** The chomo shadow homodimer (colored white and gray and displayed in cartoon) bound to a PxVxL containing peptide (green cartoon). The important residues in the peptide are showed in yellow sticks. PDBid: 1S4Z, [206].



P[MVLIRWY]V[MVLIAS][LM]

Figure 4.10: **Alignment of the verified proteins listed in table 4.5 and peptides from** [196]**.** Four out of six peptides were included in the true dataset. The remaining 2 peptides were not included due to large impact on the final regular expression. The Weblogo is made from the proteins listed in 4.5.

like that of transcription initiation factor TFIID. Since there are quite few proteins that have been identified to harbor the site, data from random phage displays obtained by [196], was included in the final regular expression.

4.5.3 Context rules and filters

Localization. Functional HP1 proteins are presumably always associated with heterochromatin in the nucleus.

Taxonomy. The HP1 is found in all eukaryotes including *S. cerevisiae* which contain a HP1-like protein Swi6p [134]. *Eukaryota*; Taxonomy ID: 2759.

Globular clash filter. There is one clash with a globular domain: In the KU70 protein from human (`SMART:Ku78`). By investigating the structure of the domain and where the functional site resides, suggests that this site is most likely accessible (results not shown).

Co-occurrence and molecular context. There are no co-occurrences of domains or other functional sites that can be associated with the PxVxL functional site. Although all proteins which contain the functional site, no surprisingly, have however domains that are associated with chromatin including Zn fingers, a SAP domain and a Bromo domain.

Table 4.5: **Verified proteins containing the PxVxL motif.**

Name	ID/Acc.nr.	Short description	Length	Dom.	PMID	Ref.	Method
CAF-1	CAFA_HUMAN	Chromatin assembly factor 1 subunit A	938	-	10549285	[158]	4
TiF1 alpha	TF1A_HUMAN	Transcription intermediary factor 1-alpha	1050	-	10562550	[163]	1
TiF1 beta	TF1B_HUMAN	Transcription intermediary factor 1-beta	835	-	10938122	[121]	4
Su(var)3-7	SUV3_DROME	Suppressor of variegation protein 3-7	1169	-	12163401	[100]	10
SETDB1	SETB_HUMAN	Histone-lysine N-methyltransferase	1291	-	11959841	[191]	6
TAFii130	T2D3_HUMAN	Transcription initiation factor TFIID subunit 4	1083	yes*	11959914	[217]	4
AF-10	Q9BMZ9	AF10	1376	-	11266362	[129]	1
Ku70	KU70_HUMAN	ATP-dependent DNA helicase II, 70 kDa subunit	608	yes	11112778	[197]	1

Table 4.6: **Statistics of the regular expressions**.

	Swiss-Prot		**Human**	
	Total hits	Hits per protein	Total hits	Hits per protein
LIG_RB	13887	0.09	4636	0.16
LIG_SID's	1533	0.01	431	0.01
LIG_CYCLIN	352194	2.3	87569	3
LIG_CtBP	22687	0.15	5937	0.2
LIG_HP1	2610	0.02	625	0.02

4.6 Attributes of the regular expressions

To get some sense of how each regular expression behave when exposed to protein sequences, searches in two different databases were performed: Swiss-Prot and a human database, see table 4.6[3]. An interesting observation is that the number of hits pr. protein in the human database increases when compared to hits pr. protein in Swiss-Prot. This may be due to human proteins are longer then the average Swiss-Prot entry. Note that these hits where not subject to filtering. See section 5.2 for discussion of predictive power of the ELM resource and demonstration of filtering.

4.7 New functional sites

During the siteseeing process and the browsing a large amount of literature, several new functional sites were identified. Some of these were annotated by a ELM consortium member or they await further analysis. These functional sites are listed in table 4.7.

[3]Searches done in Swiss-Prot release 43.0, consisting of 148516 annotated entries. A non-redundant human proteins annotated with "*Homo sapiens*" in UniProt, a total of 28814 annotated entries obtained from http://www.ebi.ac.uk/proteome/HUMAN/download.html. LIG_SID's is the sum of hits from the three regular expressions representing the SID functional site.

Table 4.7: **Functional sites identified in the literature during siteseeing.**

Name	Description	Ref.	Status	ELM ID
EEVD	TRP binding motif	[69]	annotated	LIG_TPR
PxDLS	CtBP binding motif	[188]	annotated	LIG_CtBP
YPWM	PBX binding motif	[195]	annotated	LIG_HOMEOBOX
FxxxWxxL	MDM2 binding motif	[211]	annotated	LIG_MDM2
YXXQ	STAT3 binding motif	[151]	not annotated	
HPQ	Streptavidin binding motif	[119]	not annotated	
AB	U1 snRPN 70K binding motif	[116]	not annotated	
LDSF	SAGA binding motif	[144]	not annotated	
S100B	S100B binding motif	[95]	not annotated	
LxL	MAP kinase binding site	[90]	not annotated	
MEFS	MEF2 binding site	[82]	not annotated	

Chapter 5

Discussion

During this work I have annotated five functional sites in the ELM database. In this process several issues were encountered and are discussed below. The current status of the ELM resource and future challenges are also discussed.

5.1 The Siteseeing process

Collecting instances. The siteseeing process starts with a seed paper, often a publication describing the structure of a functional site bound to recognition module or a good review. These seed publications guide the siteseer further into the literature. Ideally all references which describes a particular functional site are collected, but this can be time consuming and not trivial, since the literature can vary in quantity and focus. Performing text searches in PubMed and retrieving useful information is, to some extent, an intuitive process. To make a successful search, the query should be a combination of good keywords which describes the wanted information. Defining these keywords is the often the hardest part, but becomes easier with experience. The same can be said about doing good Google searches.

Handling the information. When collecting instances and storing it, in the form of pdf documents, sequence files and alignments, requires some form of structure and discipline. Making a local database is convenient but not necessary. There are two different strategies in siteseeing and annotation of the data: continuous or batch site-

seeing/annotation. Continuous siteseeing means that collecting data progresses slowly over time, and allow parallel analysis of several functional sites. While batch siteseeing is more intensive and focuses only one functional site at the time. In this thesis, the siteseeing process have been continuous and slow while the annotation into the ELM database have been more batch oriented, where the accumulated data have been quickly annotated.

Expanding by homology. All sequences, paralogs and orthologs could be included in the true positive dataset. A candidate motif in a homologue protein should share some degree of similarity of the motif with the verified protein. When including paralogs however, there is the risk of including false positives. Orthologs in closely related species, like human and mouse, share many of the same molecular mechanisms, but in more distant related species like human and fly, this may often not be the case. It is possible that orthologous recognition modules have diverged, which a corresponding divergence in the orthologous functional site. This will result in a poorer regular expression, which may identify the putative functional site in another species, but at the expense of the regular expression being more relaxed. There is also a limited amount of time a siteseer can use in siteseeing. Collecting orthologs of e.g. 20 verified instances of a functional site is a laborious task. This is primarily due to the large workload required and the difficulty in deciding which proteins are orthologous and which are not. A standardized automated system for expanding the true positive dataset by orthology which uses criteria such as trusted taxonomic ranges and fixed cutoff values would be very useful.

Evolving databases. Every protein listed as verified instances in the tables in Chapter 4 have a Swiss-Prot or TrEMBL id or accession number. Since these protein databases are continuously evolving and are subject to change, some of these id or accession numbers are already or may become invalid. This may be a problem since the annotated information may have been outdated and therefore useless. A possible solution is to make sure that the annotated instances in the ELM resource are automatically compared with the respective databases.

Data from viruses. Several pathogenic organisms have evolved functional sites in their proteins to interact with recognition modules in the cell, often with the intention to manipulate the cell in achieving something favorable for the virus. Such proteins have been included.

Evaluating experimental evidence. This is probably one of the most difficult parts in annotating functional sites into the ELM database. One would expect that all data in a publication have been subject to thorough examination of the authors and referees. There are, however, several scenarios where the data in a publication could be considered as questionable. Consider the following hypothetical example: a publication is investigating an interaction between two proteins and suspects that this interaction is mediated by a functional site in one of the proteins. A yeast-two hybrid system confirm the interaction. Several truncated mutants are produced in order to narrow down to a responsible subsequence. A single subsequence in one of the truncated proteins reveals a consensus motif for a functional site which is well known to mediate this protein-protein interaction. Therefor the publication concludes that this protein-protein interaction is mediated by this functional site. Should this be considered a verified functional site? This depends on the impact the publication has on the true dataset and the final regular expression. If a publication is considered as containing weak evidence, the information is only included if the publication has no or small impact on the regular expression. A large impact would be to alter a very important position e.g. the completely conserved leucine in the RxL motif, to be more relaxed e.g. to also allow the unrelated amino acid residue glutamate. Take the LxCxE functional site as an example: 25 proteins have been confirmed to contain the LxCxE functional site, among these 23 have the consensus LxCxE, while one has LxCxD and one has IxCxE. The impact of these two divergent motifs is quite large but are included due to chemical similarity and the regular expression `[LI].C.[DE]` do not overpredict too severely. It is the consequences for including weak data that should be considered. In this thesis data from almost all collected publications are included. If a publication suggests that a functional site is present without experimental support e.g. by altering the motif, this is not included.

Some studies uses viral vectors such as random page displays when investigating a functional site e.g. [183]. Undoubtedly useful, some of these data may be artificial and may not represent an in vivo situation. Data from experiments such as phage displays are only included when there are few instances of "real" proteins.

Making the regular expression. The regular expression is probably one of the most important results of the siteseeing process. There are some different views in the ELM consortium of what the regular expression should reflect: the verified instances or the molecular properties of the functional site. The regular expressions in this thesis primarily reflect the verified instances, and not always the biochemical properties of the positions. The main reason for this is that functional sites overpredict, and including none observed data will sometimes contribute to higher overprediction. In making regular expressions there are two different ways in including none observed data: excluding and including rules. Excluding rules like "not proline" in a helix, will enhance the performance of the pattern. But "leucine is observed, so why not isoleucine?" is inclusive, and will make the pattern more relaxed. A siteseers dilemma is to weight the predictive power by the pattern versus the molecular properties of the positions.

Take the functional site PxDLS as an example, see section 4.4. The functional site has been observed in 30 proteins. In the first position 28 of have a proline, but two have glycine. Should this first position be a `[P]` or a `[PG]`? The properties of proline and glycine are quite different, but should this, alone, be enough to ignore two publications which shows that glycine can indeed occur at this position? In this thesis all publications describing a functional site occurrence have been included in the true dataset and hence influence the regular expression. In this example there may be indeed a preference for having a proline at this position (or may just be biased). This knowledge are lost in the regular expression where proline and glycine are equal weighted. A more sophisticated detection method, like PSSM, may thus perform better then regular expressions. Adding to this problem is the fact that all instances of a particular functional site may have different affinities for its corresponding recognition module, which is reflected in the diversity of the sequence.

A regular expression should be under continuous development in regard to the

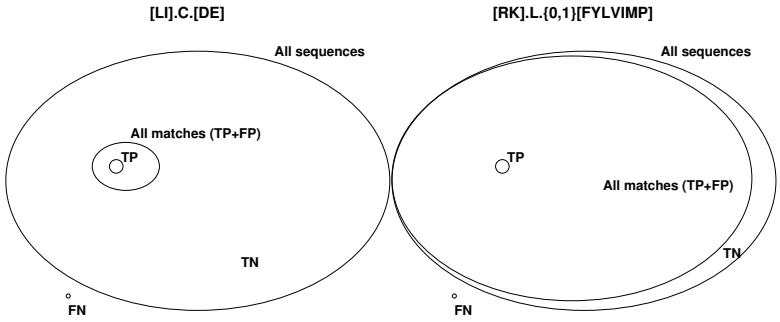
Table 5.1: **Development of the final regular expression for the RxL functional site.**

	Hits pr. human protein in Swiss-Prot	Fraction of TP from Fig. 4.6
`R.L`	3,2	16/20
`[RK].L`	6,3	20/20
`[RK].L.{0,1}[FYLIVMP]`	3,3	20/20

amount of data collected and advances reported in the scientific literature. A good example of the development of a regular expression is the RxL functional site (see Table 5.1). The regular expression started out as `R.L` (the name of the motif is RxL). Further investigation in the literature and alignment showed that the regular expression: `[RK].L` was more representative. More related structural information showed that a hydrophobic residue after the `[RK].L` motif also contributed to the interaction [135], and was thereby included in the final regular expression, `[RK].L.{0,1}[FYLIVMP]`. Often the consensus sequence in the literature (here RxL), only reflect a subset of data, and sequence and data analysis is important in making the pattern represent all verified instances.

Aspects of regular expressions. The major strength of regular expressions is that they are fast and easy to generate manually. This means that an expert can easily describe a functional site with a regular expression. Another positive thing about regular expressions is that they are easy to read and interpret in terms of biological concepts like essential residues or biochemical nature of the functional site or in the cognate recognition module. The major drawback is that they are binary in nature; either a pattern matches or it does not. This nature of patterns makes it difficult to rate and score the matches in order to find sequence similarities. It is very important that a regular expression is made carefully in respect to the rules that are applied. *Therefore the data which a regular expression is derived from, should be of high quality*.

In bioinformatics it is common to talk about the diagnostic power of a sequence analysis method. This diagnostic power is divided into two measurements: specificity and sensitivity. The most specific search returns **only** true matches. The most sensitive search returns **all** true positives. It is desirable to have high values for both (for



	Meaning	
TP	True and detected	Recovery of proteins in true dataset
FP	False and detected	Recovery of false proteins
TN	False and not detected	Loss of false proteins
FN	True and not detected	Loss of proteins in true dataset

Figure 5.1: **A graphical representation of how the regular expressions `[LI].C.[DE]` and `[RK].L.{0,1}[FYLVIMP]` behaves.** The TP are most of the true positives listed in table 4.1 and 4.3. The FN are true proteins from table 4.1 which are lost due to missing annotations in Swiss-Prot. The TN are all other sequences which does not match the regular expression.

a more extensive overview, see [94]). Calculating the values of sensitivity and specificity implies that the numbers of TP, FP, FN and TN are known (see Fig 5.1). When investigating a well known protein family this may be determined, but is far more difficult when the proteins are unrelated and share little sequence similarity. Take the regular expression `[LI].C.[DE]` from the functional site LxCxE as an example. From table 4.6, this regular expression has 13887 matches in Swiss-Prot. Approximately 20 of these are true positives (from table 4.1). Does this mean that the remaining 13867, are false positives? Probably not, since one would expect that all proteins which have the functional site are not experimentally verified yet. Since the regular expression is derived from, and matches all true positives, one would expect that the regular expression has a very high sensitivity (detects all true proteins) and have a low specificity (the true proteins are a minor fraction of the returned results). Sensitivity of the LxCxE site: TP/(TP+FN) = 20/(20+0) = 1. Specificity: TP/(TP+FP) = 20/(20+(13887-20) = 0,001. One would expect that most regular expressions describing functional sites behaves in this manner, but there are exceptions. One of these exceptions is the WRPW functional site (`ELM:LIG_WRPW_1`) which is a Groucho ligand [66]. Sensitivity = 33/(33+0) = 1. Specificity = 33/(33+21) = 0,6. Data taken from [180].

5.2 Aspects of functional sites

Aspects of functional sites. Functional sites are involved in many different types of molecular and biological processes, so it is difficult to make general statements regarding all functional sites. The ELM consortium have divided functional sites into four different classes: ligand, modification, cleavage and target sites. The ligands, as seen in this thesis, are important in mediating protein-protein interactions. Protein-protein interactions can be divided into two categories: globular domain-globular domain interactions and functional site-globular domain interactions. Functional site-globular domain interactions are thought to be short lived and transient, in contrast to more long lived protein complexes which are held together with globular domain-globular domain interactions. The five functional sites in this thesis mediates interactions between proteins to perform something to the nearby surroundings, either this is to increase the time spent in the proximity of cyclin/CDK hetrodimer or to recruit proteins with deacetylases chromatin and repress transcription. These ligand sites may allow the rapid development of novel protein-protein interactions and hence increase the connectivity between proteins.

One would think that all proteins are subject to one or more post-translational modifications during their lifetime, and therefore contain modification functional sites. Modification sites are obviously a prerequisite for the attachment of another molecule to a protein. The abundance of different functional sites which acts as modification sites, are reflected in the sheer number of different types of post-translational modifications which are known. At present, more than 200 different types of post-translational modifications are known including: phosphorylation, acetylation, glycosylation, methylation, sulfation and ubiquitination. It is also predicted that each human gene, on average, may produce three functionally different proteins as the result of post-translational modifications [24]. An abundant post-translational modification is phosphorylation, which 30% of all proteins are thought to be subject to during their functional life cycle [65]. Modification sites are regulatory determinants allowing a more dynamic behavior of proteins and expanding the functional space of amino acids and ultimately proteins.

Ligand and modification sites cooperate, in some cases, to define a molecular prop-

erty of a protein. This is seen in the p53 protein, see Figure 1.4. p53 has several ligand functional sites that interacts directly or indirectly with an enzyme and presumably increases the modification level of the protein. These ligand sites are here called substrate recognition sites. Substrate recognition sites in p53 are: ELM:LIG_CYCLIN (phosphorylation by CDK), ELM:LIG_MDM2 (ubiquitination by MDM2), LIG_p300 (acetylation by p300) and LIG_CBP (acetylation by CBP). Some other substrate recognition sites not found in p53 but in several other proteins include LIG_MAPK (MAP Kinase recognition site) [25], ELM:LIG_SH3 and ELM:LIG_SH2 (Src Homology 2 & 3 site) [192]. These substrate recognition sites may be a common mechanism to determine which proteins are modified and by who.

Many functional sites are also regulated in the same manner as domains by post-translational modifications, which is not surprising since important functions reside in them. Two examples of regulated functional sites is the acetylation of a nearby lysine residue of the PxDLS site [234] (see section 4.4.2) and the competing functional sites in the C-terminal region in p53 (see Figure 1.4). By looking at this figure p53 has stunning many functional sites. Since functional sites are recognized by domains, the number of different functional sites can not exceed the number of different globular domains. But instead the frequency of functional sites are thought to greatly outnumber the frequency of globular domains in individual proteins.

How do functional sites arise? As mentioned earlier, functional sites constitutes a major class of functional units, as seen with Src and p53 (see Fig. 1.3 and 1.4). Functional sites are a diverse group and may originate by different evolutionary mechanisms. Since functional sites are so short and occur mostly in non-globular regions, it may be possible that a handful of single point mutations at specific positions may lead to the spontaneous formation of a functional site. The prerequisites for a functional site to arise can be: 1) the presence of a globular domain which is able to act as a recognition module. 2) the host protein share some common cellular context as the protein containing the recognition module. 3) the cell benefits the functional site-recognition module interaction.

Other functional sites which are associated with an co-occurrence may share the

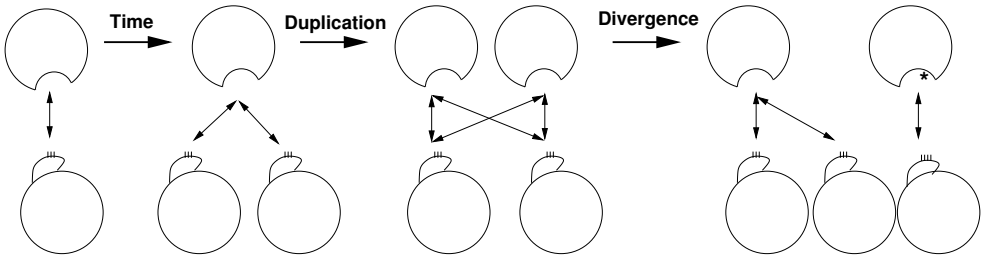


Figure 5.2: **A possible scenario for the development of a functional site.** First, on the far left, the origin of a recognition module which is able to interact with a functional site in a host protein. This mechanism is kept and over time and a new functional site spontaneously arise in a similar protein. After a duplication event of the recognition module, result in two proteins which are able to interact with the functional site containing proteins. Due to divergence, the two recognition modules develop into two slightly different modules, recognizing slightly different functional sites.

same evolutionary events as the co-occurrence. One example of this is the functional site YPWM, which is always found together with a homeo domain (SMART:HOX). The role of this functional site is to interact with another homeo domain found in the PBX protein, to increase the DNA binding affinity of the protein containing the functional site [202]. It may be possible that this functional site share, to some extent, the same evolutionary history of the homeo domain or may simply be a “satellite” part of the homeo domain.

All proteins recognizing functional sites in this thesis exist in more then one variant in human cells. This is the result of gene duplication that are though to be quite common, especially in humans compared to yeast, worm and fly [120]. This gives rise to the possibility that there may be a slight difference in selectivity of the recognition modules which is reflected in the diversity of subsequences which can be seen in Chapter 4. This possible difference in selectivity may be the result of divergence of the recognition modules which also forces a change in the cognate functional site. This is highly speculative, but interesting prospect. See Fig 5.2, for a schematic overview of a possible evolutionary scenario.

Changes in non-globular regions are more prone to genetic change and are more likely to be tolerated over time, in contrast to domains, due to their "structureless" and presumably passive nature [36]. These non-globular regions in proteins may act as a “dead subsequences” or a kind of evolutionary buffer-zone, where changes are

tolerated and a novel modification site, a protein-protein interaction site or a domain insertion may arise. It should also be mentioned that these non-globular regions may not have any apparent function, like containing functional sites, but instead act as flexible regions which are needed to separate different protein units in space. For example separate the two functional sites `ELM:MOD_CDK` and `ELM:LIG_CYCLIN`, which require a region of 12 amino acids between them to be functional [204].

What is the expected frequency of occurrence of a functional site? Consider the following hypothetical protein: A protein composed of 500 amino acids and several domains and non-globular regions. ~ 300 amino acids defines domains and the remaining ~200 amino acids reside in non-globular regions. What is the expected frequency for a functional site to occur by pure chance in the non-globular region? Take the LxCxE functional site as an example[1]: calculating the expected frequency of regular expression `[LI].C.[DE]`. Calculation is done by using the observed frequency of amino acids in *Drosophila melanogaster* [2]. The expected frequency of the pattern is $3x10^{-4}$. The expected frequency for the 200 amino acids containing the LxCxE site is therefore = $3x10^{-4}$ x 196 (since the regular expression can occur at 196 different positions) = 0,058. This leads to an interesting question: How many Drosophila proteins can we expect to contain the LxCxE functional site? 17237 proteins in a non-redundant Drosophila proteome[3] with a presumed average length of 300, results in: $3x10^{-4}$ x (17237 x 296) = 1530 proteins. Interestingly a search with the regular expression in Drosophila proteome reveals 2780 hits with the regular expression. Since the observed frequency is much higher then the expected frequency, the regular expression `[LI].C.[DE]` is not just a combination of amino acid residues, but has indeed a biological significance.

Another example of a regular expression with biological significance is a homeo domain ligand (`ELM:LIG_HOMEOBOX`). The regular expression is `[FY][DEP]WM` and have an expected frequency of $2{,}5x10^{-7}$. The expected frequency of occurrence in the Drosophila dataset is 1,46. While the observed value is 65. By this approach

[1] Here a questionable assumption is being made: that all amino acids occur uniformly in the protein sequence. There may be additional reasons for a cystine of flanked by a hydrophobic amino acids in the N-terminal and an acidic amino acid C-terminally. E.g. in a particular globular domain.

[2] Data from: http://www.ebi.ac.uk/proteome/DROME/structure/drome_7227_amino.html

[3] Obtained from: ftp://ftp.ebi.ac.uk/pub/databases/SPproteomes//fasta_files/proteomes/7227.FASTAC

it may be possible to discover new functional sites based upon the difference in expected frequency versus the observed frequency. Unfortunately however, most regular expressions describing functional sites does not show this difference in observed and expected values.

Network of the five functional sites. All collected proteins in this thesis were used to make a network with Biolayout [62]. This thesis focuses on a limited number of nuclear processes in particular transcriptional repression and may not be surprising, but instead encouraging, to see that different functional sites involved in separate protein-protein interactions, form a interconnecting network. There are several interconnecting proteins which brings one network together with another network. One of these proteins is the CtIP protein that connects the Rb and CtBP network, see figure 5.2. This is already known [148], but by constructing these kind of networks novel mechanisms and relations may be discovered. These interconnecting proteins may show that functional sites are able to bring functional networks together. As mentioned, ligand functional sites allow a rapid mechanism for not only to increase the connectivity between single proteins, but whole networks. There are six protein-protein interactions in the lower picture in figure 5.2: Three globular domain-globular domain interactions (E2F-DP, E2F-Rb and CtBP-CtBP) and three functional site-globular domain interactions (Rb-CtIP, CtIP-CtBP and CtBP-HDAC). This may suggest that functional site-domain interactions may be as common as globular domain-globular domain interactions.

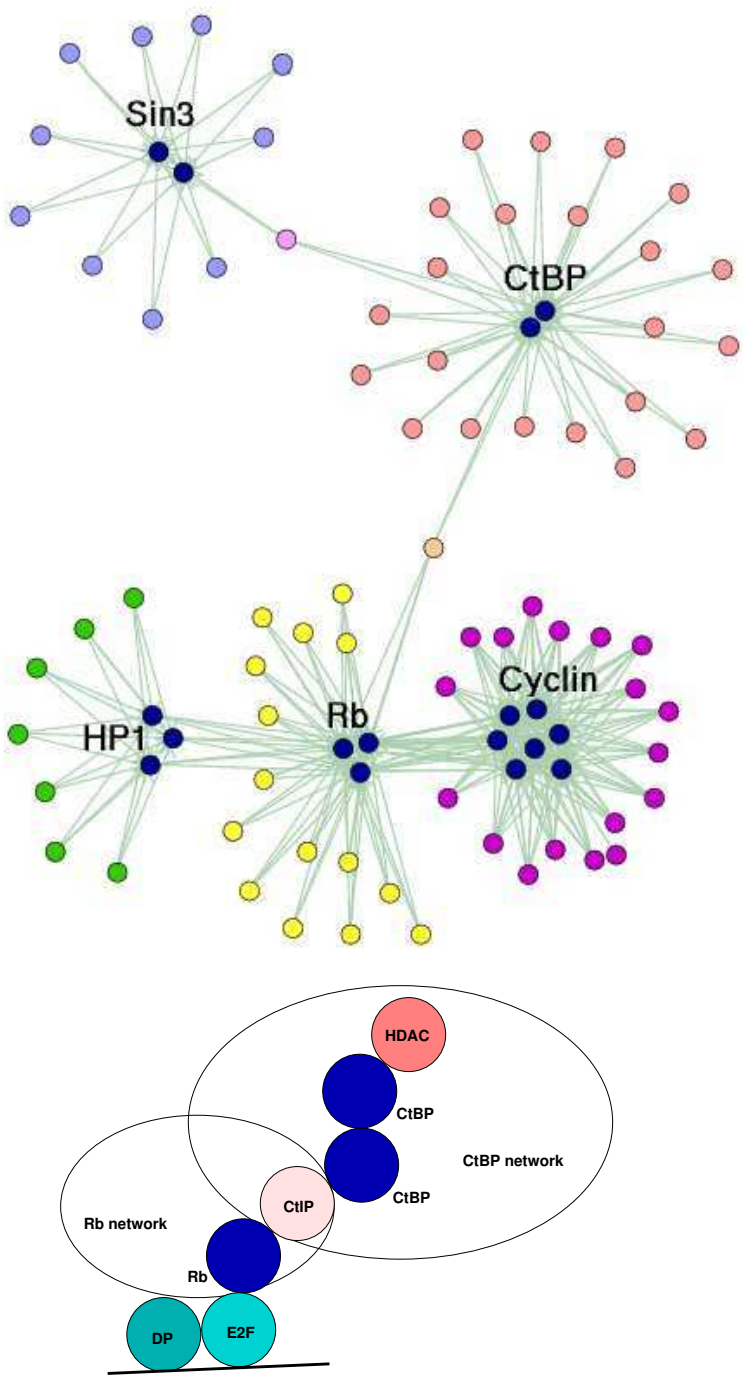


Figure 5.3: **Network of all proteins in this thesis**. Top: picture showing all interactions which involves the five functional sites analyzed here. All proteins that has a recognition module are colored blue. All proteins that contain a functional site are colored as followed: RxL=purple, LxCxE=yellow, PxVxL=green, PxDLS=red and SID=light blue. Bottom: the interconnecting protein CtIP which brings the Rb and CtBP network together via two different functional sites.

5.3 The ELM resource

This is a small summary of the current status and future challenges for the ELM resource from a siteseers perspective.

Current status. At present, there are 91 functional sites represented as 116 different patterns. The ELM resource represent the largest collection of data regarding functional sites, and there are additional ~30 functional sites which awaits siteseeing. The usage of the resource is shown in Figure 5.2.

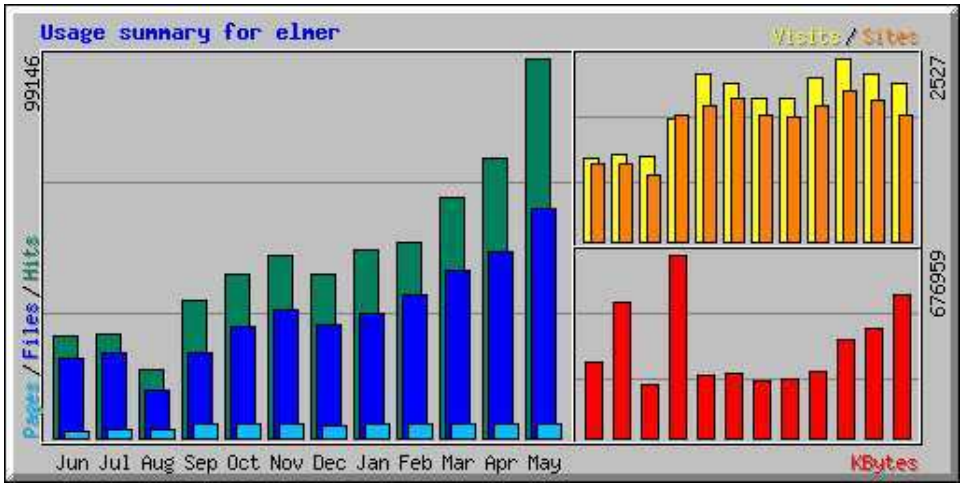


Figure 5.4: **The usage for the ELM resource.** The development of monthly number of hits for the ELM resource (http://elm.eu.org/usage.html).

Does the ELM resource work ? To what extent does the ELM resource reflect the real occurrence of functional sites in a protein? To investigate this question, a well studied protein with many experimentally verified functional sites was selected: the p53 protein from human (see section 1.1.1 for more details). p53 was subjected to superficial siteseeing, and the results are shown in Figure 1.4. Ten out of 36 experimentally verified functional sites, or 27%, are predicted by the ELM resource (seven of these are phosphorylation sites in phosphoELM, http://phospho.elm.eu.org/). This shows that many different functional sites await analysis and annotation into the ELM resource. It should be noted that all functional sites may not have been registered in this survey.

Table 5.2: **Filtering removes many candidate proteins.** Swiss-Prot 40.41 (121515 sequences) where subjected to searches with regular expressions from three functional sites and subsequently filtering of the hits. Context filters are: taxonomic distribution and sub-cellular localization. All three functional sites are found in the nucleus of metazoan species. Data from [180].

	Total Hits	Taxonomy	Subcellular localization	Fold reduction
LIG_WRPW	54	42	34	1,5
LIG_RB	6127	2784	484	12
LIG_NRBOX	44902	19963	2003	22

In section 1.2.1, the concept of filtering was explained. Table 5.3 shows that applying context information indeed removes a lot of candidate proteins. It should be mentioned that presumably some true positives are removed as well due to several factors: missing annotation in Swiss-Prot or the regular expression does not match all true positives.

Predictive power. The predictive power of the ELM resource ranges from very poor to very good, depending on the annotated functional site and its regular expression. An example of a site with very poor predictive value, is a ligand site to the PDZ domain, which has a higher frequency of occurrence then the amino acid trypthophan. Another site which overpredicts massively is the RxL functional site described in section 4.3 and have 3 hits pr. human protein, see table 4.6. On the other hand some functional sites, like the ligand to the Groucho protein, has a very high degree of confidence and most hits are to be considered true [180]. This does not mean the respective siteseers do a better job then the other, but rather that some functional sites are intrinsically very difficult to describe in respect to a regular expression. This means that most matches shown are more likely to be false positives than true matches. ELM predictions should therefore not be treated as factual findings, but instead provide useful information in experimentation. Hopefully further improvements of different filters will strengthen credibility of predictions. In addition to predict functional sites in proteins, ELM is a resource where users can browse and read about a functional site in their interest. This makes the ELM resource a valuable library for retrieving information regarding functional sites and guiding a user further into the scientific literature.

Table 5.3: **Recent ELM instances reported in the scientific literature.**

Functional site	Protein	Detected by ELM?	Ref.
LxCxE	WDW RepA	yes	[79]
RxL	Axin	yes	[111]
	Orc6	yes	[226]
PxDLS	Smad6	yes	[128]

Recent ELM instances. New verified ELM instances are frequently reported in the literature. To investigate how many new ELM instances have been reported since this work was considered finished (autumn 2003), are listed in table 5.2. The proteins in table 5.2 have not been influencing the annotated regular expression for each respective functional site. Interestingly, all four proteins are detected by the ELM resource.

5.4 Conclusions

Five functional sites were annotated into the ELM resource according to aims **i)**, **ii)** and **iii)**. Aim **iv)** is listed in table 4.6.

Functional sites is a diverse and functionally important category of protein units. This work has been a small contribution in making a needed database which can give more attention to these protein units.

Many questions remains to further investigated regarding functional sites. Some of these may be the evolution of functional sites and their cognate recognition modules. The observation that many ligand functional sites act as substrate recognition sites in post-translational modifications is also very interesting.

5.5 Future work

Many functional sites await annotation in the ELM resource as seen in table 4.6. When annotating this kind of information, development and further advances in the literature requires that the annotated information is continuously updated. This poses a major challenge for the ELM resource. Although filtering removes predictions by many orders of magnitudes, not surprisingly, overprediction is still a problem. The main reason for this is that some regular expressions are extremely short and unspecific. This simply means that some functional sites are very short, but are indeed recognized by proteins in the cell. Obviously the cell knows how to recognize these sites and may include information like a more detailed sub-cellular localization and timing the expression of relevant proteins. But we only have a protein sequence and some contextual information. Exploring and understanding the contextual information further will hopefully lead to the implementation of more detailed rules and filters. Some functional sites may simply be composed of one residue e.g. cleavage sites in proteins recognized by trypsin, which cleaves exclusively C-terminal to arginine and lysine residues [167]. Doing prediction of such a functional site is meaningless, but not if detailed contextual information are associated with the site such as: co-occurrence of functional units, expression data or molecular accessibility in terms of disorder. Maybe a more detailed sub-cellular localization information or protein complex information could provide applicable context filters. More detailed sub-cellular localizations in the nucleus could be the nucleolus, nuclear lamina, PML body or PcG body [200].

To enhance the predictive power of the ELM resource several filters are considered to be implemented. Most of these filters are structural filters, meaning they use structural information and sequence analysis the determine if a predicted functional site is accessible and hence functional. One of these filters is GlobPlot, which measures globularity and disorder of protein sequences [130]. A potential filter is applying observed relationships between protein units which are associated with a functional site, e.g. the `ELM:LIG_CYCLIN` and `ELM:MOD_CDK` are associated with one another as a logical rule, see section 4.3.3. Another example of a co-occurrence is the functional site `ELM:LIG_HOMEOBOX` which is always observed with a homeo domain (`SMART:HOX`) in several Hox proteins [202]. Another possible filter is to include protein expression data

in different cell types e.g. micro array data or annotated information from databases e.g. GXD (Mouse Gene Expression Database) [89], to discriminate between relevant predictions. For example: protein A, interacts with proteins B and C, but these two proteins: A and B, are not found simultaneously in a cell.

Context information about proteins, as applied in the ELM resource, could also be used in other online resources, like protein-protein interaction databases. Many of these databases, including MINT [231] and IntAct [87], stores binary interaction data. Although very useful in investigating a single protein-protein interaction, these resources does to a lesser extent, provide an actual picture of a biological situation.

Patterns is the most used method in detecting or searching for functional sites in protein sequences, but there exists alternatives. Meta-MEME is software tool based upon HMM models, where the focus is on strengthening statistical methods in detecting short sequences [77]. Meta-MEME builds a HMM profile based upon a MEME output format [20]. Meta-MEME is not used here, but could provide a better detection method than patterns, especially in regard to scoring hits. Another alternative are PSSM [86], and its application in ELM is being investigated. Due to emphasis on context information and not sequence detection methods, this in not investigated further in this thesis.

Currently there are nomenclature of four different classes of functional sites in the resource: ligand (LIG), modification (MOD), cleavage (CLV) and target (TRG, which is actually a ligand). This has show itself to be to simplistic. Several additional classes should be added like the isomerization (ISO & ISO_MOD) and ligand modified (LIG_MOD). An example of an ISO_MOD, is the propyl isomerase PIN1, which recognizes a phosphorylated serine/threonine and isomesmerizes the downstream proline (pS/T-P) in for instance p53, see Figure 5.3. A LIG_MOD could be the 14-3-3 proteins which recognizes and binds a phosphorylated serine/threonine and neighboring sequences [182]. One could even imagine a CLV_MOD, where a subsequence is tagged for cleavage. The problem with an event which follow after a modification, are that modification information is not implemented in standard sequence formats. One possibility is to make additional rules to an existing format like FASTA. For example a modified residue is marked with parenthesis and an additional letter or word is given to

the residue, like an acetylated lysine in the sequence ASKL then becomes AS(acK)L, or a phosphorylated serine in the same sequence, A(pS)KL.

Hopefully some experimental studies will be performed in the future based upon the predictions of the ELM resource and the functional sites described in this thesis.

Bibliography

[1] R. Aasland, C. Abrams, C. Ampe, L.J. Ball, M.T. Bedford, G. Cesareni, M. Gimona, J.H. Hurley, T. Jarchau, V.P. Lehto, M.A. Lemmon, R. Linding, B.J. Mayer, M. Nagai, M. Sudol, U. Walter, and S.J. Winder. Normalization of nomenclature for peptide motifs as ligands of modular protein domains. *FEBS Lett*, 513(1):141–4, 2002.

[2] R. Aasland and A.F. Stewart. The chromo shadow domain, a second chromo domain in heterochromatin-binding protein 1, HP1. *Nucleic Acids Res*, 23(16):3168–74, 1995.

[3] P.D. Adams, X. Li, W.R. Sellers, K.B. Baker, X. Leng, J.W. Harper, Y. Taya, and W.G. Kaelin, Jr. Retinoblastoma protein contains a C-terminal motif that targets it for phosphorylation by cyclin-cdk complexes. *Mol Cell Biol*, 19(2):1068–80, 1999.

[4] P.D. Adams, W.R. Sellers, S.K. Sharma, A.D. Wu, C.M. Nalin, and W.G. Kaelin, Jr. Identification of a cyclin-cdk2 recognition motif present in substrates and p21-like cyclin-dependent kinase inhibitors. *Mol Cell Biol*, 16(12):6623–33, 1996.

[5] J. Ahringer. NuRD and SIN3 histone deacetylase complexes in development. *Trends Genet*, 16(8):351–6, 2000.

[6] SF Altschul, W Gish, W Miller, EW Myers, and DJ Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–10, Oct 1990.

[7] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–10, 1990.

[8] H. Andersson, F. Kappeler, and H.P. Hauri. Protein targeting to endoplasmic reticulum by dilysine signals involves direct retention in addition to retrieval. *J Biol Chem*, 274(21):15080–4, 1999.

[9] E Appella and CW Anderson. Post-translational modifications and activation of p53 by genotoxic stresses. *Eur J Biochem*, 268(10):2764–72, May 2001.

[10] R. Apweiler, T.K. Attwood, A. Bairoch, A. Bateman, E. Birney, M. Biswas, P. Bucher, L. Cerutti, F. Corpet, M.D. Croning, R. Durbin, L. Falquet, W. Fleischmann, J. Gouzy, H. Hermjakob, N. Hulo, I. Jonassen, D. Kahn, A. Kanapin, Y. Karavidopoulou, R. Lopez, B. Marx, N.J. Mulder, T.M. Oinn, M. Pagni,

F. Servant, C.J. Sigrist, and E.M. Zdobnov. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res*, 29(1):37–40, 2001.

[11] Rolf Apweiler, Amos Bairoch, Cathy H Wu, Winona C Barker, Brigitte Boeckmann, Serenella Ferro, Elisabeth Gasteiger, Hongzhan Huang, Rodrigo Lopez, Michele Magrane, Maria J Martin, Darren A Natale, Claire O'Donovan, Nicole Redaschi, and Lai-Su L Yeh. UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res*, 32 Database issue:D115–9, Jan 2004.

[12] Michelle R Arkin and James A Wells. Small-molecule inhibitors of protein-protein interactions: progressing towards the dream. *Nat Rev Drug Discov*, 3(4):301–17, Apr 2004.

[13] M.N. Aronson, A.D. Meyer, J. Gyorgyey, L. Katul, H.J. Vetten, B. Gronenborn, and T. Timchenko. Clink, a nanovirus-encoded protein, binds both pRB and SKP1. *J Virol*, 74(7):2967–72, 2000.

[14] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25(1):25–9, 2000.

[15] M Ashburner and R Drysdale. FlyBase–the Drosophila genetic database. *Development*, 120(7):2077–9, Jul 1994.

[16] Terri K Attwood. The PRINTS database: a resource for identification of protein families. *Brief Bioinform*, 3(3):252–63, Sep 2002.

[17] Jose L Avalos, Ivana Celic, Shabazz Muhammad, Michael S Cosgrove, Jef D Boeke, and Cynthia Wolberger. Structure of a Sir2 enzyme bound to an acetylated p53 peptide. *Mol Cell*, 10(3):523–35, Sep 2002.

[18] S.N. Avedisov, I. Krasnoselskaya, M. Mortin, and B.J. Thomas. Roughex mediates G(1) arrest through a physical association with cyclin A. *Mol Cell Biol*, 20(21):8220–9, 2000.

[19] D.E. Ayer, C.D. Laherty, Q.A. Lawrence, A.P. Armstrong, and R.N. Eisenman. Mad proteins contain a dominant transcription repression domain. *Mol Cell Biol*, 16(10):5772–81, 1996.

[20] T.L. Bailey and M. Gribskov. Methods and statistics for combining motif match scores. *J Comput Biol*, 5(2):211–21, 1998.

[21] A Bairoch and R Apweiler. The SWISS-PROT protein sequence data bank and its new supplement TREMBL. *Nucleic Acids Res*, 24(1):21–5, Jan 1996.

[22] A Bairoch and B Boeckmann. The SWISS-PROT protein sequence data bank. *Nucleic Acids Res*, 19 Suppl:2247–9, Apr 1991.

[23] A. Bairoch, P. Bucher, and K. Hofmann. The PROSITE database, its status in 1997. *Nucleic Acids Res*, 25(1):217–21, 1997.

[24] RE Banks, MJ Dunn, DF Hochstrasser, JC Sanchez, W Blackstock, DJ Pappin, and PJ Selby. Proteomics: new perspectives, new biomedical opportunities. *Lancet*, 356(9243):1749–56, Nov 2000.

[25] AJ Bardwell, LJ Flatauer, K Matsukuma, J Thorner, and L Bardwell. A conserved docking site in MEKs mediates high-affinity binding to MAP kinases and cooperates with a scaffold protein to enhance signal transmission. *J Biol Chem*, 276(13):10374–86, Mar 2001.

[26] WC Barker, LT Hunt, DG George, LS Yeh, HR Chen, MC Blomquist, EI Seibel-Ross, A Elzanowski, JK Bair, and DA Ferrick. Protein sequence database of the protein identification resource (PIR). *Protein Seq Data Anal*, 1(1):43–98, 1987.

[27] A. Bateman, E. Birney, L. Cerruti, R. Durbin, L. Etwiller, S.R. Eddy, S. Griffiths-Jones, K.L. Howe, M. Marshall, and E.L. Sonnhammer. The Pfam protein families database. *Nucleic Acids Res*, 30(1):276–80, 2002.

[28] Alex Bateman and Daniel H Haft. HMM-based databases in InterPro. *Brief Bioinform*, 3(3):236–45, Sep 2002.

[29] HM Berman, J Westbrook, Z Feng, G Gilliland, TN Bhat, H Weissig, IN Shindyalov, and PE Bourne. The Protein Data Bank. *Nucleic Acids Res*, 28(1):235–42, Jan 2000.

[30] N. Blom, S. Gammeltoft, and S. Brunak. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J Mol Biol*, 294(5):1351–62, 1999.

[31] B. Bollag, C. Prins, E.L. Snyder, and R.J. Frisque. Association of UNP, a ubiquitin-specific protease, with the pocket proteins pRb, p107 and p130. *Oncogene*, 20(39):5533–7, 2001.

[32] F. Bonetto, M. Fanciulli, T. Battista, A. De Luca, P. Russo, T. Bruno, R. De Angelis, M. Di Padova, A. Giordano, A. Felsani, and M.G. Paggi. Interaction between the pRb2/p130 C-terminal domain and the N-terminal portion of cyclin D3. *J Cell Biochem*, 75(4):698–709, 1999.

[33] Barbara Brannetti and Manuela Helmer-Citterich. iSPOT: A web tool to infer the interaction specificity of families of protein modules. *Nucleic Acids Res*, 31(13):3709–11, Jul 2003.

[34] M. Brannon, J.D. Brown, R. Bates, D. Kimelman, and R.T. Moon. XCtBP is a XTcf-3 co-repressor with roles throughout Xenopus development. *Development*, 126(14):3159–70, 1999.

[35] S.V. Brasher, B.O. Smith, R.H. Fogh, D. Nietlispach, A. Thiru, P.R. Nielsen, R.W. Broadhurst, L.J. Ball, N.V. Murzina, and E.D. Laue. The structure of mouse HP1 suggests a unique mode of single peptide recognition by the shadow chromo domain dimer. *EMBO J*, 19(7):1587–97, 2000.

[36] C.J. Brown, S. Takayama, A.M. Campen, P. Vise, T.W. Marshall, C.J. Oldfield, C.J. Williams, and A.K. Dunker. Evolutionary rate heterogeneity in proteins with long disordered regions. *J Mol Evol*, 55(1):104–10, 2002.

[37] K. Brubaker, S.M. Cowley, K. Huang, L. Loo, G.S. Yochum, D.E. Ayer, R.N. Eisenman, and I. Radhakrishnan. Solution structure of the interacting domains of the Mad-Sin3 complex: implications for recruitment of a chromatin-modifying complex. *Cell*, 103(4):655–65, 2000.

[38] P. Bucher and A. Bairoch. A generalized profile syntax for biomolecular sequence motifs and its function in automatic sequence interpretation. *Proc Int Conf Intell Syst Mol Biol*, 2:53–61, 1994.

[39] I.M. Buyse, G. Shao, and S. Huang. The retinoblastoma protein binds to RIZ, a zinc-finger protein that shares an epitope with the adenovirus E1A protein. *Proc Natl Acad Sci U S A*, 92(10):4467–71, 1995.

[40] S. Chellappan, V.B. Kraus, B. Kroger, K. Munger, P.M. Howley, W.C. Phelps, and J.R. Nevins. Adenovirus E1A, simian virus 40 tumor antigen, and human papillomavirus E7 protein share the capacity to disrupt the interaction between transcription factor E2F and the retinoblastoma gene product. *Proc Natl Acad Sci U S A*, 89(10):4549–53, 1992.

[41] C.F. Chen, Y. Chen, K. Dai, P.L. Chen, D.J. Riley, and W.H. Lee. A new member of the hsp90 family of molecular chaperones interacts with the retinoblastoma protein during mitosis and after heat shock. *Mol Cell Biol*, 16(9):4691–9, 1996.

[42] J. Chen, P. Saha, S. Kornbluth, B.D. Dynlacht, and A. Dutta. Cyclin-binding motifs are essential for the function of p21CIP1. *Mol Cell Biol*, 16(9):4673–82, 1996.

[43] Patrick Chene. Inhibition of the p53-MDM2 interaction: targeting a protein-protein interface. *Mol Cancer Res*, 2(1):20–8, Jan 2004.

[44] D. Choubey and P. Lengyel. Binding of an interferon-inducible protein (p202) to the retinoblastoma protein. *J Biol Chem*, 270(11):6134–40, 1995.

[45] F. Corpet, F. Servant, J. Gouzy, and D. Kahn. ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Res*, 28(1):267–9, 2000.

[46] Ian G Cowell, Rebecca Aucott, Shantha K Mahadevaiah, Paul S Burgoyne, Neville Huskisson, Silvia Bongiorni, Giorgio Prantera, Laura Fanti, Sergio Pimpinelli, Rong Wu, David M Gilbert, Wei Shi, Reinald Fundele, Harris Morrison, Peter Jeppesen, and Prim B Singh. Heterochromatin, HP1 and methylation at lysine 9 of histone H3 in animals. *Chromosoma*, 111(1):22–36, Mar 2002.

[47] L Crawford. The 53,000-dalton cellular protein and its role in transformation. *Int Rev Exp Pathol*, 25:1–50, 1983.

[48] P. Criqui-Filipe, C. Ducret, S.M. Maira, and B. Wasylyk. Net, a negative Ras-switchable TCF, contains a second inhibition domain, the CID, that mediates repression through interactions with CtBP and de-acetylation. *EMBO J*, 18(12):3392–403, 1999.

[49] A. Dahiya, M.R. Gavin, R.X. Luo, and D.C. Dean. Role of the LXCXE binding site in Rb function. *Mol Cell Biol*, 20(18):6799–805, 2000.

[50] D. Defeo-Jones, P.S. Huang, R.E. Jones, K.M. Haskell, G.A. Vuocolo, M.G. Hanobik, H.E. Huber, and A. Oliff. Cloning of cDNAs for cellular proteins that bind to the retinoblastoma gene product. *Nature*, 352(6332):251–4, 1991.

[51] J. DeGregori, T. Kowalik, and J.R. Nevins. Cellular targets for activation by the E2F1 transcription factor include DNA synthesis- and G1/S-regulatory genes. *Mol Cell Biol*, 15(8):4215–24, 1995.

[52] S. Deltour, S. Pinte, C. Guerardel, B. Wasylyk, and D. Leprince. The human candidate tumor suppressor gene HIC1 recruits CtBP through a degenerate GLDLSKK motif. *Mol Cell Biol*, 22(13):4890–901, 2002.

[53] Dean J Derbyshire, Balaku P Basu, Louise C Serpell, Woo S Joo, Takayasu Date, Kuniyoshi Iwabuchi, and Aidan J Doherty. Crystal structure of human 53BP1 BRCT domains bound to p53 tumour suppressor. *EMBO J*, 21(14):3863–72, Jul 2002.

[54] D. Dornan, H. Shimizu, L. Burch, A.J. Smith, and T.R. Hupp. The proline repeat domain of p53 binds directly to the transcriptional coactivator p300 and allosterically controls DNA-dependent acetylation of p53. *Mol Cell Biol*, 23(23):8846–61, 2003.

[55] B.D. Dynlacht, K. Moberg, J.A. Lees, E. Harlow, and L. Zhu. Specific regulation of E2F family members by cyclin-dependent kinases. *Mol Cell Biol*, 17(7):3867–75, 1997.

[56] S.R. Eddy. Profile hidden Markov models. *Bioinformatics*, 14(9):755–63, 1998.

[57] B Eisenhaber, P Bork, and F Eisenhaber. Prediction of potential GPI-modification sites in proprotein sequences. *J Mol Biol*, 292(3):741–58, Sep 1999.

[58] WS el Deiry, SE Kern, JA Pietenpol, KW Kinzler, and B Vogelstein. Definition of a consensus binding site for p53. *Nat Genet*, 1(1):45–9, Apr 1992.

[59] C.J. Elferink, N.L. Ge, and A. Levine. Maximal aryl hydrocarbon receptor activity depends on an interaction with the retinoblastoma protein. *Mol Pharmacol*, 59(4):664–73, 2001.

[60] V. Ellenrieder, J.S. Zhang, J. Kaczynski, and R. Urrutia. Signaling disrupts mSin3A binding to the Mad1-like Sin3-interacting domain of TIEG2, an Sp1-like repressor. *EMBO J*, 21(10):2451–60, 2002.

[61] O Emanuelsson, H Nielsen, S Brunak, and G von Heijne. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J Mol Biol*, 300(4):1005–16, Jul 2000.

[62] AJ Enright and CA Ouzounis. BioLayout–an automatic graph layout algorithm for similarity visualization. *Bioinformatics*, 17(9):853–4, Sep 2001.

[63] I. Fernandes, Y. Bastien, T. Wai, K. Nygard, R. Lin, O. Cormier, H.S. Lee, F. Eng, N.R. Bertos, N. Pelletier, S. Mader, V.K. Han, X.J. Yang, and J.H. White. Ligand-Dependent Nuclear Receptor Corepressor LCoR Functions by Histone Deacetylase-Dependent and -Independent Mechanisms. *Mol Cell*, 11(1):139–50, 2003.

[64] D. Fesquet, J.C. Labbe, J. Derancourt, J.P. Capony, S. Galas, F. Girard, T. Lorca, J. Shuttleworth, M. Doree, and J.C. Cavadore. The MO15 gene encodes the catalytic subunit of a protein kinase that activates cdc2 and other cyclin-dependent kinases (CDKs) through phosphorylation of Thr161 and its homologues. *EMBO J*, 12(8):3111–21, 1993.

[65] Scott B Ficarro, Mark L McCleland, P Todd Stukenberg, Daniel J Burke, Mark M Ross, Jeffrey Shabanowitz, Donald F Hunt, and Forest M White. Phosphoproteome analysis by mass spectrometry and its application to Saccharomyces cerevisiae. *Nat Biotechnol*, 20(3):301–5, Mar 2002.

[66] AL Fisher, S Ohsako, and M Caudy. The WRPW motif of the hairy-related basic helix-loop-helix repressor proteins acts as a 4-amino-acid transcription repression and protein-protein interaction domain. *Mol Cell Biol*, 16(6):2670–7, Jun 1996.

[67] T.C. Fleischer, U.J. Yun, and D.E. Ayer. Identification and characterization of three new components of the mSin3A corepressor complex. *Mol Cell Biol*, 23(10):3456–67, 2003.

[68] R.Y. Forng and C.D. Atreya. Mutations in the retinoblastoma protein-binding LXCXE motif of rubella virus putative replicase affect virus replication. *J Gen Virol*, 80 (Pt 2):327–32, 1999.

[69] B.C. Freeman, M.P. Myers, R. Schumacher, and R.I. Morimoto. Identification of a regulatory motif in Hsp70 that affects ATPase activity, substrate binding and interaction with HDJ-1. *EMBO J*, 14(10):2281–92, 1995.

[70] Jeffery E.F. Friedl. *Mastering Regular Expressions - Powerful Techniques for Perl and Other Tools*. O'Reilly, 1997.

[71] T. Furusawa, H. Moribe, H. Kondoh, and Y. Higashi. Identification of CtBP1 and CtBP2 as corepressors of zinc finger-homeodomain factor deltaEF1. *Mol Cell Biol*, 19(12):8581–90, 1999.

[72] W Gilbert, SJ de Souza, and M Long. Origin of genes. *Proc Natl Acad Sci U S A*, 94(15):7698–703, Jul 1997.

[73] T. Goda, T. Ishii, N. Nakajo, N. Sagata, and H. Kobayashi. The RRASK motif in Xenopus cyclin B2 is required for the substrate recognition of Cdc25C by the cyclin B-Cdc2 complex. *J Biol Chem*, 278(21):19032–7, 2003.

[74] Or Gozani, Philip Karuman, David R Jones, Dmitri Ivanov, James Cha, Alexey A Lugovskoy, Cheryl L Baird, Hong Zhu, Seth J Field, Stephen L Lessnick, Jennifer Villasenor, Bharat Mehrotra, Jian Chen, Vikram R Rao, Joan S Brugge, Colin G Ferguson, Bernard Payrastre, David G Myszka, Lewis C Cantley, Gerhard Wagner, Nullin Divecha, Glenn D Prestwich, and Junying Yuan. The PHD finger of the chromatin-associated protein ING2 functions as a nuclear phosphoinositide receptor. *Cell*, 114(1):99–111, Jul 2003.

[75] G. Grafi, R.J. Burnett, T. Helentjaris, B.A. Larkins, J.A. DeCaprio, W.R. Sellers, and W.G. Kaelin, Jr. A maize cDNA encoding a member of the retinoblastoma protein family: involvement in endoreduplication. *Proc Natl Acad Sci U S A*, 93(17):8962–7, 1996.

[76] M Gribskov, AD McLachlan, and D Eisenberg. Interaction between Smad-interacting protein-1 and the corepressor C-terminal binding protein is dispensable for transcriptional repression of E-cadherin. *Proc Natl Acad Sci U S A*, 84(13):4355–8, Jul 1987.

[77] W.N. Grundy, T.L. Bailey, C.P. Elkan, and M.E. Baker. Meta-MEME: motif-based hidden Markov models of protein families. *Comput Appl Biosci*, 13(4):397–406, 1997.

[78] Jian Gu, Lidong Zhang, Stephen G Swisher, Jinsong Liu, Jack A Roth, and Bingliang Fang. Induction of p53-regulated genes in lung cancer cells: implications of the mechanism for adenoviral p53-mediated apoptosis. *Oncogene*, 23(6):1300–7, Feb 2004.

[79] Crisanto Gutierrez, Elena Ramirez-Parra, M Mar Castellano, Andres P Sanz-Burgos, Alejandro Luque, and Riccardo Missich. Geminivirus DNA replication and cell cycle interactions. *Vet Microbiol*, 98(2):111–9, Feb 2004.

[80] DH Haft, BJ Loftus, DL Richardson, F Yang, JA Eisen, IT Paulsen, and O White. TIGRFAMs: a protein family resource for the functional identification of proteins. *Nucleic Acids Res*, 29(1):41–3, Jan 2001.

[81] C. Hall, D.M. Nelson, X. Ye, K. Baker, J.A. DeCaprio, S. Seeholzer, M. Lipinski, and P.D. Adams. HIRA, the human homologue of yeast Hir1p and Hir2p, is a novel cyclin-cdk2 substrate whose expression blocks S-phase progression. *Mol Cell Biol*, 21(5):1854–65, 2001.

[82] Aidong Han, Fan Pan, James C Stroud, Hong-Duk Youn, Jun O Liu, and Lin Chen. Sequence-specific recruitment of transcriptional co-repressor Cabin1 by myocyte enhancer factor-2. *Nature*, 422(6933):730–4, Apr 2003.

[83] JE Hansen, O Lund, N Tolstrup, AA Gooley, KL Williams, and S Brunak. NetOglyc: prediction of mucin type O-glycosylation sites based on sequence context and surface accessibility. *Glycoconj J*, 15(2):115–30, Feb 1998.

[84] K. Hansen, T. Farkas, J. Lukas, K. Holm, L. Ronnstrand, and J. Bartek. Phosphorylation-dependent and -independent functions of p130 cooperate to evoke a sustained G1 block. *EMBO J*, 20(3):422–32, 2001.

[85] Y. Hashimoto, K. Kohri, Y. Kaneko, H. Morisaki, T. Kato, K. Ikeda, and M. Nakanishi. Critical role for the 310 helix region of p57(Kip2) in cyclin-dependent kinase 2 inhibition and growth suppression. *J Biol Chem*, 273(26):16544–50, 1998.

[86] S Henikoff and JG Henikoff. Embedding strategies for effective use of information from multiple sequence alignments. *Protein Sci*, 6(3):698–705, Mar 1997.

[87] Henning Hermjakob, Luisa Montecchi-Palazzi, Chris Lewington, Sugath Mudali, Samuel Kerrien, Sandra Orchard, Martin Vingron, Bernd Roechert, Peter Roepstorff, Alfonso Valencia, Hanah Margalit, John Armstrong, Amos Bairoch, Gianni Cesareni, David Sherman, and Rolf Apweiler. IntAct: an open source molecular interaction database. *Nucleic Acids Res*, 32 Database issue:D452–5, Jan 2004.

[88] L. Hertel, S. Rolle, M. De Andrea, B. Azzimonti, R. Osello, G. Gribaudo, M. Gariglio, and S. Landolfo. The retinoblastoma protein is an essential mediator that links the interferon-inducible 204 gene to cell-cycle regulation. *Oncogene*, 19(32):3598–608, 2000.

[89] David P Hill, Dale A Begley, Jacqueline H Finger, Terry F Hayamizu, Ingeborg J McCright, Constance M Smith, Jon S Beal, Lori E Corbani, Judith A Blake, Janan T Eppig, James A Kadin, Joel E Richardson, and Martin Ringwald. The mouse Gene Expression Database (GXD): updates and enhancements. *Nucleic Acids Res*, 32 Database issue:D568–71, Jan 2004.

[90] Justine M Hill, Hema Vaidyanathan, Joe W Ramos, Mark H Ginsberg, and Milton H Werner. Recognition of ERK MAP kinase by PEA-15 reveals a common docking site within the death domain and death effector domain. *EMBO J*, 21(23):6494–504, Dec 2002.

[91] M Hollstein, D Sidransky, B Vogelstein, and CC Harris. p53 mutations in human cancers. *Science*, 253(5015):49–53, Jul 1991.

[92] SR Hubbard. Src autoinhibition: let us count the ways. *Nat Struct Biol*, 6(8):711–4, Aug 1999.

[93] P.J. Hurlin, C. Queva, and R.N. Eisenman. Mnt: a novel Max-interacting protein and Myc antagonist. *Curr Top Microbiol Immunol*, 224:115–21, 1997.

[94] William R. Taylor Ingvar Eidhammer, Inge Jonassen. *Protein Bioinformatics: An Algorithmic Approach to Sequence and Structure Analysis*. John Wiley & Sons, Ltd, 2004.

[95] Keith G Inman, Ruiqing Yang, Richard R Rustandi, Kristine E Miller, Donna M Baldisseri, and David J Weber. Solution NMR structure of S100B bound to the high-affinity target peptide TRTK-12. *J Mol Biol*, 324(5):1003–14, Dec 2002.

[96] A Ito, CH Lai, X Zhao, S Saito, MH Hamilton, E Appella, and TP Yao. p300/CBP-mediated p53 acetylation is commonly induced by p53-activating agents and inhibited by MDM2. *EMBO J*, 20(6):1331–40, Mar 2001.

[97] K. Izutsu, M. Kurokawa, Y. Imai, M. Ichikawa, T. Asai, K. Maki, K. Mitani, and H. Hirai. The t(3;21) fusion product, AML1/Evi-1 blocks AML1-induced transactivation by recruiting CtBP. *Oncogene*, 21(17):2695–703, 2002.

[98] James R Jabbur, Amy D Tabor, Xiaodong Cheng, Hua Wang, Motonari Uesugi, Guillermina Lozano, and Wei Zhang. Mdm-2 binding and TAF(II)31 recruitment is regulated by hydrogen bond disruption between the p53 residues Thr18 and Asp21. *Oncogene*, 21(46):7100–13, Oct 2002.

[99] S.A. Jacobs, S.D. Taverna, Y. Zhang, S.D. Briggs, J. Li, J.C. Eissenberg, C.D. Allis, and S. Khorasanizadeh. Specificity of the HP1 chromo domain for the methylated N-terminus of histone H3. *EMBO J*, 20(18):5232–41, 2001.

[100] Y. Jaquet, M. Delattre, A. Spierer, and P. Spierer. Functional dissection of the Drosophila modifier of variegation Su(var)3-7. *Development*, 129(17):3975–82, 2002.

[101] Iddo Friedberg Jeff Chang, Brad Chapman. Biopython tutorial and cookbook. Available online, http://www.bioinformatics.org/bradstuff/bp/tut/Tutorial.pdf.

[102] J. Kaczynski, J.S. Zhang, V. Ellenrieder, A. Conley, T. Duenes, H. Kester, B. van Der Burg, and R. Urrutia. The Sp1-like protein BTEB3 inhibits transcription via the basic transcription element box by interacting with mSin3A and HDAC-1 co-repressors and competing with Sp1. *J Biol Chem*, 276(39):36749–56, 2001.

[103] D. Kadosh and K. Struhl. Repression by Ume6 involves recruitment of a complex containing Sin3 corepressor and Rpd3 histone deacetylase to target promoters. *Cell*, 89(3):365–71, 1997.

[104] R.F. Kalejta, J.T. Bechtel, and T. Shenk. Human Cytomegalovirus pp71 Stimulates Cell Cycle Progression by Inducing the Proteasome-Dependent Degradation of the Retinoblastoma Family of Tumor Suppressors. *Mol Cell Biol*, 23(6):1885–95, 2003.

[105] Minoru Kanehisa, Susumu Goto, Shuichi Kawashima, and Akihiro Nakaya. The KEGG databases at GenomeNet. *Nucleic Acids Res*, 30(1):42–6, Jan 2002.

[106] T. Kaneko, T. Katoh, S. Sato, A. Nakamura, E. Asamizu, and S. Tabata. Structural analysis of Arabidopsis thaliana chromosome 3. II. Sequence features of the 4,251,695 bp regions covered by 90 P1, TAC and BAC clones. *DNA Res*, 7(3):217–21, 2000.

[107] S.G. Katz, A.B. Cantor, and S.H. Orkin. Interaction between FOG-1 and the corepressor C-terminal binding protein is dispensable for normal erythropoiesis in vivo. *Mol Cell Biol*, 22(9):3121–8, 2002.

[108] S.A. Keller, Y. Mao, P. Struffi, C. Margulies, C.E. Yurk, A.R. Anderson, R.L. Amey, S. Moore, J.M. Ebels, K. Foley, M. Corado, and D.N. Arnosti. dCtBP-dependent and -independent repression activities of the Drosophila Knirps protein. *Mol Cell Biol*, 20(19):7247–58, 2000.

[109] HY Kim, BY Ahn, and Y Cho. Structural basis for the inactivation of retinoblastoma tumor suppressor by SV40 large T antigen. *EMBO J*, 20(1-2):295–304, Jan 2001.

[110] H.Y. Kim, B.Y. Ahn, and Y. Cho. Structural basis for the inactivation of retinoblastoma tumor suppressor by SV40 large T antigen. *EMBO J*, 20(1-2):295–304, 2001.

[111] Sung Il Kim, Chun Shik Park, Mi Su Lee, Min Seong Kwon, Eek hoon Jho, and Woo Keun Song. Cyclin-dependent kinase 2 regulates the interaction of Axin with beta-catenin. *Biochem Biophys Res Commun*, 317(2):478–83, Apr 2004.

[112] E.S. Knudsen and J.Y. Wang. Differential regulation of retinoblastoma protein function by specific Cdk phosphorylation sites. *J Biol Chem*, 271(14):8313–20, 1996.

[113] J. Koipally and K. Georgopoulos. Ikaros interactions with CtBP reveal a repression mechanism that is independent of histone deacetylase activity. *J Biol Chem*, 275(26):19594–602, 2000.

[114] MM Krem and E Di Cera. Molecular markers of serine protease evolution. *EMBO J*, 20(12):3036–45, Jun 2001.

[115] V. Kumar, J.E. Carlson, K.A. Ohgi, T.A. Edwards, D.W. Rose, C.R. Escalante, M.G. Rosenfeld, and A.K. Aggarwal. Transcription corepressor CtBP is an NAD(+)-regulated dehydrogenase. *Mol Cell*, 10(4):857–69, 2002.

[116] E Labourier, MD Adams, and DC Rio. Modulation of P-element pre-mRNA splicing by a direct interaction between PSI and U1 snRNP 70K protein. *Mol Cell*, 8(2):363–73, Aug 2001.

[117] A. Lai, B.K. Kennedy, D.A. Barbie, N.R. Bertos, X.J. Yang, M.C. Theberge, S.C. Tsai, E. Seto, Y. Zhang, A. Kuzmichev, W.S. Lane, D. Reinberg, E. Harlow, and P.E. Branton. RBP1 recruits the mSIN3-histone deacetylase complex to the pocket of retinoblastoma tumor suppressor family proteins found in limited discrete regions of the nucleus at growth arrest. *Mol Cell Biol*, 21(8):2918–32, 2001.

[118] ND Lakin and SP Jackson. Regulation of p53 in response to DNA damage. *Oncogene*, 18(53):7644–55, Dec 1999.

[119] T. Lamla and V.A. Erdmann. Searching sequence space for high-affinity binding peptides using ribosome display. *J Mol Biol*, 329(2):381–8, 2003.

[120] ES Lander, LM Linton, B Birren, C Nusbaum, MC Zody, J Baldwin, K Devon, K Dewar, M Doyle, W FitzHugh, R Funke, D Gage, K Harris, A Heaford,

J Howland, L Kann, J Lehoczky, R LeVine, P McEwan, K McKernan, J Meldrim, JP Mesirov, C Miranda, W Morris, J Naylor, C Raymond, M Rosetti, R Santos, A Sheridan, C Sougnez, N Stange-Thomann, N Stojanovic, A Subramanian, D Wyman, J Rogers, J Sulston, R Ainscough, S Beck, D Bentley, J Burton, C Clee, N Carter, A Coulson, R Deadman, P Deloukas, A Dunham, I Dunham, R Durbin, L French, D Grafham, S Gregory, T Hubbard, S Humphray, A Hunt, M Jones, C Lloyd, A McMurray, L Matthews, S Mercer, S Milne, JC Mullikin, A Mungall, R Plumb, M Ross, R Shownkeen, S Sims, RH Waterston, RK Wilson, LW Hillier, JD McPherson, MA Marra, ER Mardis, LA Fulton, AT Chinwalla, KH Pepin, WR Gish, SL Chissoe, MC Wendl, KD Delehaunty, TL Miner, A Delehaunty, JB Kramer, LL Cook, RS Fulton, DL Johnson, PJ Minx, SW Clifton, T Hawkins, E Branscomb, P Predki, P Richardson, S Wenning, T Slezak, N Doggett, JF Cheng, A Olsen, S Lucas, C Elkin, E Uberbacher, M Frazier, RA Gibbs, DM Muzny, SE Scherer, JB Bouck, EJ Sodergren, KC Worley, CM Rives, JH Gorrell, ML Metzker, SL Naylor, RS Kucherlapati, DL Nelson, GM Weinstock, Y Sakaki, A Fujiyama, M Hattori, T Yada, A Toyoda, T Itoh, C Kawagoe, H Watanabe, Y Totoki, T Taylor, J Weissenbach, R Heilig, W Saurin, F Artiguenave, P Brottier, T Bruls, E Pelletier, C Robert, P Wincker, DR Smith, L Doucette-Stamm, M Rubenfield, K Weinstock, HM Lee, J Dubois, A Rosenthal, M Platzer, G Nyakatura, S Taudien, A Rump, H Yang, J Yu, J Wang, G Huang, J Gu, L Hood, L Rowen, A Madan, S Qin, RW Davis, NA Federspiel, AP Abola, MJ Proctor, RM Myers, J Schmutz, M Dickson, J Grimwood, DR Cox, MV Olson, R Kaul, C Raymond, N Shimizu, K Kawasaki, S Minoshima, GA Evans, M Athanasiou, R Schultz, BA Roe, F Chen, H Pan, J Ramser, H Lehrach, R Reinhardt, WR McCombie, M de la Bastide, N Dedhia, H BlÃ¶cker, K Hornischer, G Nordsiek, R Agarwala, L Aravind, JA Bailey, A Bateman, S Batzoglou, E Birney, P Bork, DG Brown, CB Burge, L Cerutti, HC Chen, D Church, M Clamp, RR Copley, T Doerks, SR Eddy, EE Eichler, TS Furey, J Galagan, JG Gilbert, C Harmon, Y Hayashizaki, D Haussler, H Hermjakob, K Hokamp, W Jang, LS Johnson, TA Jones, S Kasif, A Kaspryzk, S Kennedy, WJ Kent, P Kitts, EV Koonin, I Korf, D Kulp, D Lancet, TM Lowe, A McLysaght, T Mikkelsen, JV Moran, N Mulder, VJ Pollara, CP Ponting, G Schuler, J Schultz, G Slater, AF Smit, E Stupka, J Szustakowski, D Thierry-Mieg, J Thierry-Mieg, L Wagner, J Wallis, R Wheeler, A Williams, YI Wolf, KH Wolfe, SP Yang, RF Yeh, F Collins, MS Guyer, J Peterson, A Felsenfeld, KA Wetterstrand, A Patrinos, MJ Morgan, J Szustakowki, P de Jong, JJ Catanese, K Osoegawa, H Shizuya, S Choi, YJ Chen, and YJ Chen. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, Feb 2001.

[121] M.S. Lechner, G.E. Begg, D.W. Speicher, and F.J. Rauscher, 3rd. Molecular determinants for targeting heterochromatin protein 1-mediated gene silencing: direct chromoshadow domain-KAP-1 corepressor interaction is essential. *Mol Cell Biol*, 20(17):6449–65, 2000.

[122] Changwook Lee, Jeong Ho Chang, Hyun Sook Lee, and Yunje Cho. Structural basis for the recognition of the E2F transactivation domain by the retinoblas-

toma tumor suppressor. *Genes Dev*, 16(24):3199–212, Dec 2002.

[123] J.O. Lee, A.A. Russo, and N.P. Pavletich. Structure of the retinoblastoma tumour-suppressor pocket domain bound to a peptide from HPV E7. *Nature*, 391(6670):859–65, 1998.

[124] X. Leng, M. Noble, P.D. Adams, J. Qin, and J.W. Harper. Reversal of growth suppression by p107 via direct phosphorylation by cyclin D1/cyclin-dependent kinase 4. *Mol Cell Biol*, 22(7):2242–54, 2002.

[125] I. Letunic, L. Goodstadt, N.J. Dickens, T. Doerks, J. Schultz, R. Mott, F. Ciccarelli, R.R. Copley, C.P. Ponting, and P. Bork. Recent improvements to the SMART domain-based sequence annotation resource. *Nucleic Acids Res*, 30(1):242–4, 2002.

[126] SH Liang and MF Clarke. A bipartite nuclear localization signal is required for p53 nuclear import regulated by a carboxyl-terminal domain. *J Biol Chem*, 274(46):32699–703, Nov 1999.

[127] X. Lin, P. Nelson, and I.H. Gelman. SSeCKS, a major protein kinase C substrate with tumor suppressor activity, regulates G(1)–>S progression by controlling the expression and cellular compartmentalization of cyclin D. *Mol Cell Biol*, 20(19):7259–72, 2000.

[128] Xia Lin, Yao-Yun Liang, Baohua Sun, Min Liang, Yujiang Shi, F Charles Brunicardi, Yang Shi, and Xin-Hua Feng. Smad6 recruits transcription corepressor CtBP to repress bone morphogenetic protein-induced transcription. *Mol Cell Biol*, 23(24):9081–93, Dec 2003.

[129] B. Linder, N. Gerlach, and H. Jackle. The Drosophila homolog of the human AF10 is an HP1-interacting suppressor of position effect variegation. *EMBO Rep*, 2(3):211–6, 2001.

[130] Rune Linding, Robert B Russell, Victor Neduva, and Toby J Gibson. GlobPlot: Exploring protein sequences for globularity and disorder. *Nucleic Acids Res*, 31(13):3701–8, Jul 2003.

[131] F. Liu, C. Rothblum-Oviatt, C.E. Ryan, and H. Piwnica-Worms. Overproduction of human Myt1 kinase induces a G2 cell cycle delay by interfering with the intracellular trafficking of Cdc2-cyclin B1 complexes. *Mol Cell Biol*, 19(7):5113–23, 1999.

[132] L. Liu, K. Saunders, C.L. Thomas, J.W. Davies, and J. Stanley. Bean yellow dwarf virus RepA, but not rep, binds to maize retinoblastoma protein, and the virus tolerates mutations in the consensus binding motif. *Virology*, 256(2):270–9, 1999.

[133] MA Lohrum, DB Woods, RL Ludwig, E BÃ¡lint, and KH Vousden. C-terminal ubiquitination of p53 contributes to nuclear export. *Mol Cell Biol*, 21(24):8521–32, Dec 2001.

[134] A Lorentz, K Ostermann, O Fleck, and H Schmidt. Switching gene swi6, involved in repression of silent mating-type loci in fission yeast, encodes a homologue of chromatin-associated proteins from Drosophila and mammals. *Gene*, 143(1):139–43, May 1994.

[135] E.D. Lowe, I. Tews, K.Y. Cheng, N.R. Brown, S. Gul, M.E. Noble, S.J. Gamblin, and L.N. Johnson. Specificity determinants of recruitment peptides bound to phospho-CDK2/cyclin A. *Biochemistry*, 41(52):15625–34, 2002.

[136] Edward D Lowe, Ivo Tews, Kin Yip Cheng, Nick R Brown, Sheraz Gul, Martin E M Noble, Steven J Gamblin, and Louise N Johnson. Specificity determinants of recruitment peptides bound to phospho-CDK2/cyclin A. *Biochemistry*, 41(52):15625–34, Dec 2002.

[137] X. Lu and H.R. Horvitz. lin-35 and lin-53, two genes that antagonize a C. elegans Ras pathway, encode proteins similar to Rb and its binding protein RbAp48. *Cell*, 95(7):981–91, 1998.

[138] M.G. Luciani, J.R. Hutchins, D. Zheleva, and T.R. Hupp. The C-terminal regulatory domain of p53 contains a functional docking site for cyclin A. *J Mol Biol*, 300(3):503–18, 2000.

[139] B. Luscher. Function and regulation of the transcription factors of the Myc/Max/Mad network. *Gene*, 277(1-2):1–14, 2001.

[140] Han-Hui Ma, Li Yang, and Bo-Liang Li. Expression, purification and in vitro N-myristoylation of human Src N-terminal region. *Sheng Wu Hua Xue Yu Sheng Wu Wu Li Xue Bao (Shanghai)*, 35(1):13–7, Jan 2003.

[141] T. Ma, B.A. Van Tine, Y. Wei, M.D. Garrett, D. Nelson, P.D. Adams, J. Wang, J. Qin, L.T. Chow, and J.W. Harper. Cell cycle-regulated phosphorylation of p220(NPAT) by cyclin E/Cdk2 in Cajal bodies promotes histone gene transcription. *Genes Dev*, 14(18):2298–313, 2000.

[142] W.R. MacLellan, G. Xiao, M. Abdellatif, and M.D. Schneider. A novel Rb- and p300-binding protein inhibits transactivation by MyoD. *Mol Cell Biol*, 20(23):8903–15, 2000.

[143] GS Martin. The hunting of the Src. *Nat Rev Mol Cell Biol*, 2(6):467–75, Jun 2001.

[144] ME Massari, PA Grant, MG Pray-Grant, SL Berger, JL Workman, and C Murre. A conserved motif present in a class of helix-loop-helix proteins activates transcription by direct recruitment of the SAGA complex. *Mol Cell*, 4(1):63–73, Jul 1999.

[145] Sebastian Maurer-Stroh, Birgit Eisenhaber, and Frank Eisenhaber. N-terminal N-myristoylation of proteins: prediction of substrate proteins from amino acid sequence. *J Mol Biol*, 317(4):541–57, Apr 2002.

[146] M McCoy, ES Stavridi, JL Waterman, AM Wieczorek, SJ Opella, and TD Halazonetis. Hydrophobic side-chain size is a determinant of the three-dimensional structure of the p53 oligomerization domain. *EMBO J*, 16(20):6230–6, Oct 1997.

[147] T.A. Melhuish and D. Wotton. The interaction of the carboxyl terminus-binding protein with the Smad corepressor TGIF is disrupted by a holoprosencephaly mutation in TGIF. *J Biol Chem*, 275(50):39762–6, 2000.

[148] A.R. Meloni, E.J. Smith, and J.R. Nevins. A mechanism for Rb/p130-mediated transcription repression involving recruitment of the CtBP corepressor. *Proc Natl Acad Sci U S A*, 96(17):9574–9, 1999.

[149] Nerissa Mendoza, Sharon Fong, Jim Marsters, Hartmut Koeppen, Ralph Schwall, and Dineli Wickramasinghe. Selective cyclin-dependent kinase 2/cyclin A antagonists that differ from ATP site inhibitors block tumor growth. *Cancer Res*, 63(5):1020–4, Mar 2003.

[150] G. Meroni, S. Cairo, G. Merla, S. Messali, R. Brent, A. Ballabio, and A. Reymond. Mlx, a new Max-like bHLHZip family member: the center stage of a novel transcription factors regulatory pathway? *Oncogene*, 19(29):3266–77, 2000.

[151] M. Minoguchi, S. Minoguchi, D. Aki, A. Joo, T. Yamamoto, T. Yumioka, T. Matsuda, and A. Yoshimura. STAP-2/BKS, an adaptor/docking protein, modulates STAT3 activation in acute-phase response through its YXXQ motif. *J Biol Chem*, 278(13):11182–9, 2003.

[152] F. Monigatti, E. Gasteiger, A. Bairoch, and E. Jung. The Sulfinator: predicting tyrosine sulfation sites in protein sequences. *Bioinformatics*, 18(5):769–70, 2002.

[153] D.O. Morgan. Cyclin-dependent kinases: engines, clocks, and microprocessors. *Annu Rev Cell Dev Biol*, 13:261–91, 1997.

[154] S.D. Morgenbesser, B.O. Williams, T. Jacks, and R.A. DePinho. p53-dependent apoptosis produced by Rb-deficiency in the developing mouse lens. *Nature*, 371(6492):72–4, 1994.

[155] Shiraz Mujtaba, Yan He, Lei Zeng, Sherry Yan, Olga Plotnikova, Olga Sachchidanand, Roberto Sanchez, Nancy J Zeleznik-Le, Ze'ev Ronai, and Ming-Ming Zhou. Structural mechanism of the bromodomain of the coactivator CBP in p53 transcriptional activation. *Mol Cell*, 13(2):251–63, Jan 2004.

[156] Stefan Muller, Andreas Ledl, and Darja Schmidt. SUMO: a regulator of gene expression and genome integrity. *Oncogene*, 23(11):1998–2008, Mar 2004.

[157] A. Murakami, S. Ishida, J. Thurlow, J.M. Revest, and C. Dickson. SOX6 binds CtBP2 to repress transcription from the Fgf-3 promoter. *Nucleic Acids Res*, 29(16):3347–55, 2001.

[158] R. Nair, P. Carter, and B. Rost. Heterochromatin dynamics in mouse cells: interaction between chromatin assembly factor 1 and HP1 proteins. *Mol Cell*, 4(4):529–40, 1999.

[159] R. Nair, P. Carter, and B. Rost. NLSdb: database of nuclear localization signals. *Nucleic Acids Res*, 31(1):397–9, 2003.

[160] Satish K Nair and Stephen K Burley. X-ray structures of Myc-Max and Mad-Max recognizing DNA. Molecular bases of regulation by proto-oncogenic transcription factors. *Cell*, 112(2):193–205, Jan 2003.

[161] K Nakai and P Horton. PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem Sci*, 24(1):34–6, Jan 1999.

[162] M. Nardini, S. Spano, C. Cericola, A. Pesce, A. Massaro, E. Millo, A. Luini, D. Corda, and M. Bolognesi. CtBP/BARS: a dual-function protein involved in transcription co-repression and Golgi membrane fission. *EMBO J*, 22(12):3122–30, 2003.

[163] A.L. Nielsen, J.A. Ortiz, J. You, M. Oulad-Abdelghani, R. Khechumian, A. Gansmuller, P. Chambon, and R. Losson. Interaction with members of the heterochromatin protein 1 (HP1) family and histone deacetylation are differentially involved in transcriptional silencing by members of the TIF1 family. *EMBO J*, 18(22):6385–95, 1999.

[164] H Nielsen, S Brunak, and G von Heijne. Machine learning approaches for the prediction of signal peptides and other protein sorting signals. *Protein Eng*, 12(1):3–9, Jan 1999.

[165] John C Obenauer, Lewis C Cantley, and Michael B Yaffe. Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res*, 31(13):3635–41, Jul 2003.

[166] A. Ohtoshi and H. Otoshi. Analysis of beta3-endonexin mutants for their ability to interact with cyclin A. *Mol Genet Genomics*, 266(4):664–71, 2001.

[167] Jesper V. Olsen, Shao-En Ong, and Matthias Mann. Trypsin cleaves exclusively C-terminal to Arginine and lysine residues. *Mol Cell Proteomics*, Mar 2004.

[168] Y. Oma, K. Nishimori, and M. Harata. The brain-specific actin-related protein ArpNalpha interacts with the transcriptional co-repressor CtBP. *Biochem Biophys Res Commun*, 301(2):521–8, 2003.

[169] T. Pawson, G.D. Gish, and P. Nash. SH2 domains, interaction modules and cellular wiring. *Trends Cell Biol*, 11(12):504–11, 2001.

[170] T. Pawson and P. Nash. Assembly of cell regulatory systems through protein interaction domains. *Science*, 300(5618):445–52, 2003.

[171] Tony Pawson. Specificity in signal transduction: from phosphotyrosine-SH2 domain interactions to complex cellular systems. *Cell*, 116(2):191–203, Jan 2004.

[172] WR Pearson and DJ Lipman. Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A*, 85(8):2444–8, Apr 1988.

[173] V. Pennaneach, I. Salles-Passador, A. Munshi, H. Brickner, K. Regazzoni, F. Dick, N. Dyson, T.T. Chen, J.Y. Wang, R. Fotedar, and A. Fotedar. The large subunit of replication factor C promotes cell survival after DNA damage in an LxCxE motif- and Rb-dependent manner. *Mol Cell*, 7(4):715–27, 2001.

[174] J. Perdomo and M. Crossley. The Ikaros family protein Eos associates with C-terminal-binding protein corepressors. *Eur J Biochem*, 269(23):5885–92, 2002.

[175] B.O. Petersen, J. Lukas, C.S. Sorensen, J. Bartek, and K. Helin. Phosphorylation of mammalian CDC6 by cyclin A/CDK2 regulates its subcellular localization. *EMBO J*, 18(2):396–410, 1999.

[176] C.P. Ponting and R.R. Russell. The natural history of protein domains. *Annu Rev Biophys Biomol Struct*, 31:45–71, 2002.

[177] G. Poortinga, M. Watanabe, and S.M. Parkhurst. Drosophila CtBP: a Hairy-interacting protein required for embryonic segmentation and hairy-mediated transcriptional repression. *EMBO J*, 17(7):2067–78, 1998.

[178] AA Postigo and DC Dean. ZEB represses transcription through interaction with the corepressor CtBP. *Proc Natl Acad Sci U S A*, 96(12):6683–8, Jun 1999.

[179] L.R. Pratt and A. Pohorille. Hydrophobic effects and modeling of biophysical aqueous solution interfaces. *Chem Rev*, 102(8):2671–92, 2002.

[180] P. Puntervoll, R. Linding, C. Gemund, S. Chabanis-Davidson, M. Mattingsdal, S. Cameron, D.M. Martin, G. Ausiello, B. Brannetti, A. Costantini, F. Ferre, V. Maselli, A. Via, G. Cesareni, F. Diella, G. Superti-Furga, L. Wyrwicz, C. Ramu, C. McGuigan, R. Gudavalli, I. Letunic, P. Bork, L. Rychlewski, B. Kuster, M. Helmer-Citterich, W.N. Hunter, R. Aasland, and T.J. Gibson. ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Res*, 31(13):3625–30, 2003.

[181] Anja Ragvin, Havard Valvatne, Sigrid Erdal, Vibeke Arskog, Katharina R Tufteland, Kamilla Breen, Anne M Yan, Anton Eberharter, Toby J Gibson, Peter B Becker, and Rein Aasland. Nucleosome binding by the bromodomain and PHD finger of the transcriptional cofactor p300. *J Mol Biol*, 337(4):773–88, Apr 2004.

[182] Michael R Roberts. 14-3-3 proteins find new partners in plant cell signalling. *Trends Plant Sci*, 8(5):218–23, May 2003.

[183] D.J. Rodi and L. Makowski. Phage-display technology–finding a needle in a vast molecular haystack. *Curr Opin Biotechnol*, 10(1):87–93, 1999.

[184] E. Rubin, S. Mittnacht, E. Villa-Moruzzi, and J.W. Ludlow. Site-specific and temporally-regulated retinoblastoma protein dephosphorylation by protein phosphatase type 1. *Oncogene*, 20(29):3776–85, 2001.

[185] RR Rustandi, DM Baldisseri, and DJ Weber. Structure of the negative regulatory domain of p53 bound to S100B(betabeta). *Nat Struct Biol*, 7(7):570–4, Jul 2000.

[186] P. Saha, Q. Eichbaum, E.D. Silberman, B.J. Mayer, and A. Dutta. p21CIP1 and Cdc25A: competition between an inhibitor and an activator of cyclin-dependent kinases. *Mol Cell Biol*, 17(8):4338–45, 1997.

[187] M. Santaguida, Q. Ding, G. Berube, M. Truscott, P. Whyte, and A. Nepveu. Phosphorylation of the CCAAT displacement protein (CDP)/Cux transcription factor by cyclin A-Cdk1 modulates its DNA binding activity in G(2). *J Biol Chem*, 276(49):45780–90, 2001.

[188] U. Schaeper, J.M. Boyd, S. Verma, E. Uhlmann, T. Subramanian, and G. Chinnadurai. Molecular cloning and characterization of a cellular phosphoprotein that interacts with a conserved C-terminal domain of adenovirus E1A involved in negative modulation of oncogenic transformation. *Proc Natl Acad Sci U S A*, 92(23):10467–71, 1995.

[189] U. Schaeper, T. Subramanian, L. Lim, J.M. Boyd, and G. Chinnadurai. Interaction between a cellular protein that binds to the C-terminal region of adenovirus E1A (CtBP) and a novel cellular protein is disrupted by E1A through a conserved PLDLS motif. *J Biol Chem*, 273(15):8549–52, 1998.

[190] TD Schneider and RM Stephens. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res*, 18(20):6097–100, Oct 1990.

[191] David C Schultz, Kasirajan Ayyanathan, Dmitri Negorev, Gerd G Maul, and Frank J Rauscher. SETDB1: a novel KAP-1-associated histone H3, lysine 9-specific methyltransferase that contributes to HP1-mediated silencing of euchromatic genes by KRAB zinc-finger proteins. *Genes Dev*, 16(8):919–32, Apr 2002.

[192] MP Scott and WT Miller. A peptide model system for processive phosphorylation by Src family kinases. *Biochemistry*, 39(47):14531–7, Nov 2000.

[193] R.G. Sewalt, M.J. Gunster, J. van der Vlag, D.P. Satijn, and A.P. Otte. C-Terminal binding protein is a transcriptional repressor that interacts with a specific class of vertebrate Polycomb proteins. *Mol Cell Biol*, 19(1):777–87, 1999.

[194] P.B. Singh, J.R. Miller, J. Pearce, R. Kothary, R.D. Burton, R. Paro, T.C. James, and S.J. Gaunt. A sequence motif found in a Drosophila heterochromatin protein is conserved in animals and plants. *Nucleic Acids Res*, 19(4):789–94, 1991.

[195] C.M. Slupsky, D.B. Sykes, G.L. Gay, and B.D. Sykes. The HoxB1 hexapeptide is a prefolded domain: implications for the Pbx1/Hox interaction. *Protein Sci*, 10(6):1244–53, 2001.

[196] J.F. Smothers and S. Henikoff. The HP1 chromo shadow domain binds a consensus peptide pentamer. *Curr Biol*, 10(1):27–30, 2000.

[197] K. Song, Y. Jung, D. Jung, and I. Lee. Human Ku70 interacts with heterochromatin protein 1alpha. *J Biol Chem*, 276(11):8321–7, 2001.

[198] EL Sonnhammer, G von Heijne, and A Krogh. A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc Int Conf Intell Syst Mol Biol*, 6:175–82, 1998.

[199] C.S. Sorensen, C. Lukas, E.R. Kramer, J.M. Peters, J. Bartek, and J. Lukas. A conserved cyclin-binding domain determines functional interplay between anaphase-promoting complex-Cdh1 and cyclin A-Cdk2 during cell cycle progression. *Mol Cell Biol*, 21(11):3692–703, 2001.

[200] David L Spector. The dynamics of chromosome organization and gene regulation. *Annu Rev Biochem*, 72:573–608, 2003.

[201] C.A. Spronk, M. Tessari, A.M. Kaan, J.F. Jansen, M. Vermeulen, H.G. Stunnenberg, and G.W. Vuister. The Mad1-Sin3B interaction involves a novel helical fold. *Nat Struct Biol*, 7(12):1100–4, 2000.

[202] Tara Sprules, Nancy Green, Mark Featherstone, and Kalle Gehring. Lock and key binding of the HOX YPWM peptide to the PBX homeodomain. *J Biol Chem*, 278(2):1053–8, Jan 2003.

[203] JM Stommel, ND Marchenko, GS Jimenez, UM Moll, TJ Hope, and GM Wahl. A leucine-rich nuclear export signal in the p53 tetramerization domain: regulation of subcellular localization and p53 activity by NES masking. *EMBO J*, 18(6):1660–72, Mar 1999.

[204] D.Y. Takeda, J.A. Wohlschlegel, and A. Dutta. A bipartite substrate recognition motif for cyclin-dependent kinases. *J Biol Chem*, 276(3):1993–7, 2001.

[205] Y. Tao, R.F. Kassatly, W.D. Cress, and J.M. Horowitz. Subunit composition determines E2F DNA-binding site specificity. *Mol Cell Biol*, 17(12):6994–7007, 1997.

[206] Abarna Thiru, Daniel Nietlispach, Helen R Mott, Mitsuru Okuwaki, Debbie Lyon, Peter R Nielsen, Miriam Hirshberg, Alain Verreault, Natalia V Murzina, and Ernest D Laue. Structural basis of HP1/PXVXL motif peptide interactions and HP1 localisation to heterochromatin. *EMBO J*, 23(3):489–99, Feb 2004.

[207] J.D. Thompson, T.J. Gibson, F. Plewniak, F. Jeanmougin, and D.G. Higgins. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res*, 25(24):4876–82, 1997.

[208] JD Thompson, DG Higgins, and TJ Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, 22(22):4673–80, Nov 1994.

[209] D. Trouche, C. Le Chalony, C. Muchardt, M. Yaniv, and T. Kouzarides. RB and hbrm cooperate to repress the activation functions of E2F1. *Proc Natl Acad Sci U S A*, 94(21):11268–73, 1997.

[210] J. Turner and M. Crossley. Cloning and characterization of mCtBP2, a co-repressor that associates with basic Kruppel-like factor and other mammalian transcriptional regulators. *EMBO J*, 17(17):5129–40, 1998.

[211] M. Uesugi and G.L. Verdine. The alpha-helical FXXPhiPhi motif in p53: TAF interaction and discrimination by MDM2. *Proc Natl Acad Sci U S A*, 96(26):14801–6, 1999.

[212] J.G. Umen and U.W. Goodenough. Control of cell division by a retinoblastoma protein homolog in Chlamydomonas. *Genes Dev*, 15(13):1652–61, 2001.

[213] Leo A van Grunsven, Christine Michiels, Tom Van de Putte, Luc Nelles, Gunther Wuytens, Kristin Verschueren, and Danny Huylebroeck. Interaction between Smad-interacting protein-1 and the corepressor C-terminal binding protein is dispensable for transcriptional repression of E-cadherin. *J Biol Chem*, 278(28):26135–45, Jul 2003.

[214] Guido van Rossum. Python tutorial. Available: http://www.python.org/doc/current/tut/tut.html.

[215] J. van Vliet, J. Turner, and M. Crossley. Human Kruppel-like factor 8: a CACCC-box binding protein that associates with CtBP and represses transcription. *Nucleic Acids Res*, 28(9):1955–62, 2000.

[216] L. Vandel, E. Nicolas, O. Vaute, R. Ferreira, S. Ait-Si-Ali, and D. Trouche. Transcriptional repression by the retinoblastoma protein through the recruitment of a histone methyltransferase. *Mol Cell Biol*, 21(19):6484–94, 2001.

[217] M.F. Vassallo and N. Tanese. Isoform-specific interaction of HP1 with human TAFII130. *Proc Natl Acad Sci U S A*, 99(9):5919–24, 2002.

[218] N. Vo, C. Fjeld, and R.H. Goodman. Acetylation of nuclear hormone receptor-interacting protein RIP140 regulates binding of the transcriptional corepressor CtBP. *Mol Cell Biol*, 21(18):6181–8, 2001.

[219] C.Y. Wang, B. Petryniak, C.B. Thompson, W.G. Kaelin, and J.M. Leiden. Regulation of the Ets-related transcription factor Elf-1 by binding to the retinoblastoma protein. *Science*, 260(5112):1330–5, 1993.

[220] G Wang, A Ma, CM Chow, D Horsley, NR Brown, IG Cowell, and PB Singh. Conservation of heterochromatin protein 1 function. *Mol Cell Biol*, 20(18):6970–83, Sep 2000.

[221] R. Wanitchakorn, G.J. Hafner, R.M. Harding, and J.L. Dale. Functional analysis of proteins encoded by banana bunchy top virus DNA-4 to -6. *J Gen Virol*, 81(Pt 1):299–306, 2000.

[222] R. Weigert, M.G. Silletta, S. Spano, G. Turacchio, C. Cericola, A. Colanzi, S. Senatore, R. Mancini, E.V. Polishchuk, M. Salmona, F. Facchiano, K.N. Burger, A. Mironov, A. Luini, and D. Corda. CtBP/BARS induces fission of Golgi membranes by acylating lysophosphatidic acid. *Nature*, 402(6760):429–33, 1999.

[223] H. Wen and S. Ao. RBP95, a novel leucine zipper protein, binds to the retinoblastoma protein. *Biochem Biophys Res Commun*, 275(1):141–8, 2000.

[224] Y. Wen, D. Nguyen, Y. Li, and Z.C. Lai. The N-terminal BTB/POZ domain and C-terminal sequences are essential for Tramtrack69 to specify cell fate in the developing Drosophila eye. *Genetics*, 156(1):195–203, 2000.

[225] D.L. Wheeler, D.M. Church, A.E. Lash, D.D. Leipe, T.L. Madden, J.U. Pontius, G.D. Schuler, L.M. Schriml, T.A. Tatusova, L. Wagner, and B.A. Rapp. Database resources of the National Center for Biotechnology Information: 2002 update. *Nucleic Acids Res*, 30(1):13–6, 2002.

[226] Gwendolyn M. Wilmes, Vincent Archambault, Richard J. Austin, Matthew D. Jacobson, Stephen P. Bell, and Frederick R. Cross. Interaction of the S-phase cyclin Clb5 with an RXL docking sequence in the initiator protein Orc6 provides an origin-localized replication control switch. *Genes Dev*, Apr 2004.

[227] J.T. Woitach, M. Zhang, C.H. Niu, and S.S. Thorgeirsson. A retinoblastoma-binding protein that affects cell-cycle control and confers transforming ability. *Nat Genet*, 19(4):371–4, 1998.

[228] D. Wotton, P.S. Knoepfler, C.D. Laherty, R.N. Eisenman, and J. Massague. The Smad transcriptional corepressor TGIF recruits mSin3. *Cell Growth Differ*, 12(9):457–63, 2001.

[229] J. Yang and S. Kornbluth. All aboard the cyclin train: subcellular trafficking of cyclins and their CDK partners. *Trends Cell Biol*, 9(6):207–10, 1999.

[230] G.S. Yochum and D.E. Ayer. Pf1, a novel PHD zinc finger protein that links the TLE corepressor to the mSin3A-histone deacetylase complex. *Mol Cell Biol*, 21(13):4110–8, 2001.

[231] Andreas Zanzoni, Luisa Montecchi-Palazzi, Michele Quondam, Gabriele Ausiello, Manuela Helmer-Citterich, and Gianni Cesareni. MINT: a Molecular INTeraction database. *FEBS Lett*, 513(1):135–40, Feb 2002.

[232] C.L. Zhang, T.A. McKinsey, J.R. Lu, and E.N. Olson. Association of COOH-terminal-binding protein (CtBP) and MEF2-interacting transcription repressor (MITR) contributes to transcriptional repression of the MEF2 transcription factor. *J Biol Chem*, 276(1):35–9, 2001.

[233] J.S. Zhang, M.C. Moncrieffe, J. Kaczynski, V. Ellenrieder, F.G. Prendergast, and R. Urrutia. A conserved alpha-helical motif mediates the interaction of Sp1-like transcriptional repressors with the corepressor mSin3A. *Mol Cell Biol*, 21(15):5041–9, 2001.

[234] Q. Zhang, H. Yao, N. Vo, and R.H. Goodman. Acetylation of adenovirus E1A regulates binding of the transcriptional corepressor CtBP. *Proc Natl Acad Sci U S A*, 97(26):14323–8, 2000.

[235] Hongwu Zheng, Han You, Xiao Zhen Zhou, Stephen A Murray, Takafumi Uchida, Gerburg Wulf, Ling Gu, Xiaoren Tang, Kun Ping Lu, and Zhi-Xiong Jim Xiao. The prolyl isomerase Pin1 is a regulator of p53 in genotoxic response. *Nature*, 419(6909):849–53, Oct 2002.

[236] JT Zilfou, WH Hoffman, M Sank, DL George, and M Murphy. The corepressor mSin3a interacts with the proline-rich domain of p53 and protects p53 from proteasome-mediated degradation. *Mol Cell Biol*, 21(12):3974–85, Jun 2001.

Appendix

Algorithm 1 getseq python code. Used for online retrieval of protein sequences.

```
#!/usr/bin/python2.2
#
# getseq [ Accession, entry name ]
#
# Author Morten Mattingsdal

import os,sys
from Bio.WWW import ExPASy

try:
ids = [sys.argv[1]]
except:
print "\n getseq [id/acc.nr]\n"
raise SystemExit

all_results = ''
for id in ids:
        results = ExPASy.get_sprot_raw(id)
        all_results = all_results + results.read()

from Bio.SwissProt import SProt
from Bio import File
s_parser = SProt.RecordParser()
s_iterator = SProt.Iterator(File.StringHandle(all_results), s_parser)
while 1:
        cur_record = s_iterator.next()
        if cur_record is None:
                break
        print ">" + cur_record.entry_name + "\n" + cur_record.sequence + "\n"
        print "length", cur_record.sequence_length
        print
```