Fachhochschul-Masterstudiengang
BIOMEDIZINISCHE INFORMATIK
4232 Hagenberg, Austria

Computational Prediction of Kinase-Substrate Interactions with Scansite 3 and the Enrichment of well-known Protein-Protein Interaction Networks with Novel and Relevant Interactors

Masterarbeit

zur Erlangung des akademischen Grades
Master of Science in Engineering

Eingereicht von

Tobias Ehrenberger, BSc

Betreuer:	Prof. Michael B Yaffe, MD PhD, MIT, USA
Begutachter:	Prof. (FH) Dipl.-Ing. Peter W Kulczycki

Juni 2012

Declaration

I hereby declare and confirm that this thesis is entirely the result of my own original work. Where other sources of information have been used, they have been indicated as such and properly acknowledged. I further declare that this or similar work has not been submitted for credit elsewhere.

Ort, Datum	Tobias Ehrenberger

Acknowledgements

I would like to thank Professor Michael Yaffe who offered me to be part of his lab and Jonathan Rameseder, who introduced me to this amazing group of people in the first place. This group of people includes all members of the Yaffe-lab (and extends to people from the Hemann-Lab) as they not only helped me whenever necessary and inspired me in many ways, but also have become amazing friends that made working there a lot of fun. Thank you for all the time we spent together and for all the rewarding conversations! It was an honour to be a part of this group!

I also want to express my gratitude to the Marshall Plan Foundation that generously supported this project and my stay in the US, and my Professors here in Austria, especially my thesis advisor Prof. Peter Kulczycki.

Finally, and most importantly, I want to thank all my other friends and my family, particularly my parents, for all their support and the little pushes they give me whenever I need them!

Tobias Ehrenberger

Computational Prediction of Kinase-Substrate Interactions with Scansite 3 and the Enrichment of well-known Protein-Protein Interaction Networks with Novel and Relevant Interactors

Kurzfassung

Lebende Zellen sind komplexe Systeme die stark von einem funktionierendem Signalübertragungsapparat abhängig sind. Alle Signaltransduktionsereignisse basieren auf der Interaktion von Proteinen mit anderen Proteinen oder anderen Molekülen. Eine der wichtigsten Arten von Interaktionen in diesem Kontext ist Phosphorylierung, ein zellulärer Prozess der von Proteinkinasen katalysiert wird. Viele Datenbanken von experimentell verifizierten Phosphorylierungs-Interaktionen sind online verfügbar, jedoch sind diese bei weitem noch nicht vollständig. Computerprogramme werden deswegen zum Schließen dieser Lücken benötigt. In dieser Arbeit wird eine neue Version eines *in silico* Kinasen-Substrat Interaktionsvorhersageprogrammes, das auf kurzen linearen Sequenzmotifen basiert, vorgestellt: Scansite 3 (*http://scansite3.mit.edu/*) hat eine neue Benutzeroberfläche, eine Menge neuer Funktionen, und beinhaltet ein Webservice welches programmatischen Zugriff erlaubt. Hier wird beschrieben wie Scansite funktioniert und wie es unterstützend eingesetzt werden kann um verschiedene wissenschaftlich relevante Fragen zu beantworten. Zusätzlich wird eine Methode vorgestellt die es erlaubt eine Menge von Genen mit einem für eine gewisse Fragestellung relevantem Interaktionsnetzwerk zu assoziieren und die wichtigsten Interaktionen aus diesem Netzwerk zu extrahieren. Um möglicherweise noch unbekannte, aber relevante, Signalwege identifizieren zu können werden Phosphorylierungsvorhersagen von Scansite herangezogen. Diese Methode wird hier dazu eingesetzt um eine algorithmische Pipeline vorzubereiten, die RNAi-Screen Ergebnisse mit dem Interaktionsnetzwerk der molekularen Antwort auf Schäden der DNS („DNA damage response“) verbindet. Da bislang noch kein Interaktionsnetzwerk dieser Art publiziert worden ist, wurde ein Netzwerk dieser Art händisch aus einer Vielzahl von Publikationen extrahiert. Dieses Netzwerk wird hier ebenso vorgestellt, da es als relevant und hilfreich für Wissenschaftler, die sich mit der DNA damage response auseinandersetzen, erachtet wird.

Schlüsselwörter: *Scansite, Kinasen, Phosphorylierung, DNA damage response, Protein-Protein Interaktionen, Price Collecting Steiner Tree*

Abstract

Living cells are complex systems that are highly dependent on a properly functioning cell signalling apparatus. All signalling events are based on the interaction of proteins with one another or with other molecules. One of the most important types of interactions is phosphorylation, a process that is catalysed by protein kinases. Many databases of experimentally verified kinase-substrate interactions are available online, but these by far do not cover all phosphorylation events that happen in living cells. Therefore, computational tools are needed to fill these gaps. Here, a new version of an *in silico* kinase-substrate interaction prediction tool that is based on short-linear sequence motifs is presented: Scansite 3 (*http://scansite3.mit.edu/*) has a new user interface, a number of new features, and includes a web service that allows computational access. This work describes how Scansite works and how it can be used to assist in answering different scientifically relevant questions. In addition, a method is presented that allows the association of a set of genes with an interaction network of interest which also reduces this network to only the most important interactions. In order to be able to identify novel signalling pathways, it includes predictions of Scansite 3 to include new interactions. Here, this strategy is applied to prepare an algorithmic pipeline that links RNAi-screen results with the DNA damage response interaction network. Since no interaction network like this has been published so far, a manually curated DNA damage response interaction network is presented here too. This network may be a useful resource for scientists that do work related to the molecular response to DNA damage.

Keywords: *Scansite, kinases, phosphorylation, DNA damage response, protein-protein interaction, Price Collecting Steiner Tree*

Contents

Chapter 1

Introduction

One of the most feared diseases of this age is cancer. It originates from cells that behave in an abnormal way and many times can be traced back to mutated genes that give rise to non-functional or misbehaving proteins which influence the regulation of many cell signalling pathways in a bad way. All signalling events in biological systems depend on the interaction of molecules with one another. Signalling in cells can be viewed on different levels: The most basic level is the level of single interactions of proteins with other proteins or molecules. For example, many transcription factors are activated by phosphorylation. Phosphorylation is one of the most important protein-protein interaction types known as it is involved in almost all cellular signal transduction events. Single interactions are important cellular events, but analysing only them does not give a lot of information about *why* (i. e. in what biological context) they happen. This question can be answered by analysing cascades of multiple interactions like this. A one-after-another cascade of interactions is referred to as a signalling pathway. Pathways are generally used to explain how signalling events are controlled in a simplified way. However, a more accurate view is that of a signalling network (Figure 1.1). Signalling in biological networks is a complex interaction of many parties, many of which collaborate in order to achieve a result, others have opposing interests (e. g. kinases and phosphatases targeting the same protein) in which case those with the higher binding affinity or the higher concentration wins. Thus, Levy et al. (2010) refer to this interaction network as a "democratic" network, comparing it to "a table around which decision-makers debate a question and respond collectively to information put to them". However, although the network-view is more accurate it comes at the cost of being very complex, which makes it incredibly hard to analyse. Years of experimental efforts gave rise to enormous databases of experimentally verified biological information, including protein-protein interaction information. Nonetheless, even these huge datasets give just a glimpse into what is happening in a cell interaction-wise. Hence, computational tools are needed that assist in determining which interactions are the most relevant to certain events in a cell. Given that the interaction data we have is most probably not complete, computational interaction prediction tools can be used to extend this dataset with further interactions that may be of potential significance.

Computational tools have been successful in many parts of research in molecular biology, be it

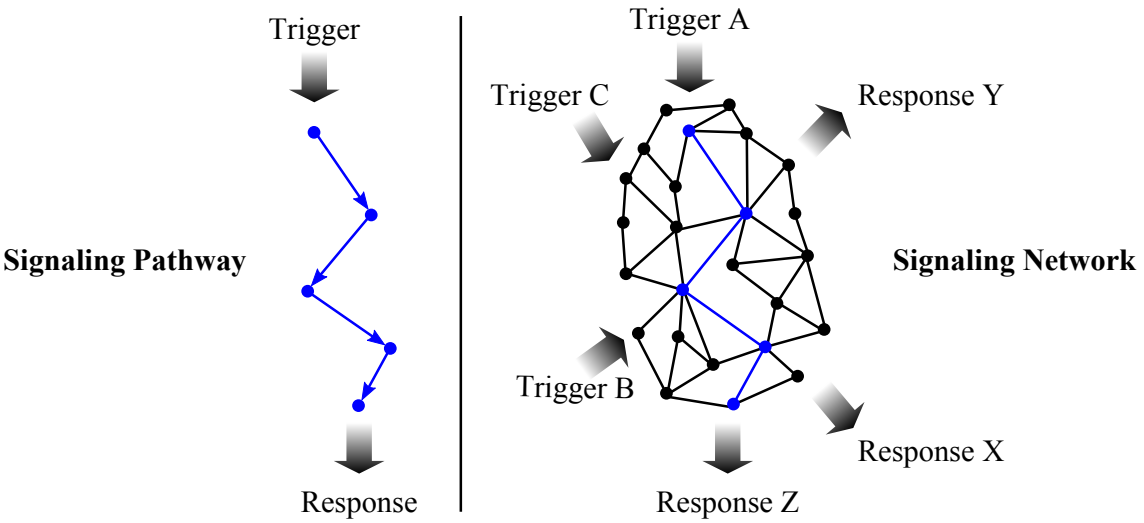


Figure 1.1: Two different views of signalling in biological systems: On the left, the traditional pathway-view that is initiated by a trigger event on one side and leads to some molecular response on the other end. The more accurate view of signalling interactions is presented on the right-hand side in a network-view of the system. The entities in the network are closely interconnected and influence each other in different ways. Traditional pathways may be embedded in this network, but are influenced by a wide variety of other interactors too. Changes to the network may be induced on different levels and yield different results. Although the directionality of the interactions in the network are not shown in the figure, feedback loops and other combinations of directed interactions add a new layer of complexity to signalling networks and make analyses a hard task. Hence, the simplified pathway-view is more widely used.

storing and managing data, or assisting in analysing datasets (visualisation of data, calculating statistics, etc.). The availability of vast amounts of data combined with the use of computational tools offers an easy way to create new hypotheses following the *low-hanging fruit strategy*: Obvious hypotheses (either available in the datasets, or promising results generated by computational tools which only need to be verified or refuted) can be seen as low-hanging fruits that are ready to be picked and examined. But applying this strategy is not always as straight-forward as it seems, especially when it comes to putting an experiment's results in context with the huge datasets that are already available and trying to identify what pieces of information are important for the question at hand.

Here, a method is described that assists in doing exactly that in the context of network biology. Given a set of genes that an experiment determined to be important, this method places them in the context of a known protein-protein interaction network. In order to omit the complex view of such a network, the network is then reduced to only the most important interactions, giving a pathway-like view of the examined genes in the context of the research (i. e. the interaction-network of interest). In order to not restrict this method to what is already known alone, predicted interactions can be included, therefore allowing to find new pathways that cannot be found by just considering experimental interaction-datasets. This method is demonstrated by associating genes identified

by an RNAi-screen with the DNA damage response protein-protein interaction network. The putative interactions that are included come from a new version of Scansite, a phosphorylation and binding-motif prediction tool that is also presented here. Scansite 3 is now available online at *http://scansite3.mit.edu/*.

In the following chapters all biological concepts that are relevant for understanding how Scansite 3 and the gene-network-association method work are described in detail. This includes the biology of protein-protein interactions in general (focusing on phosphorylation) and the biological and clinical relevance of the DNA damage response. The description of Scansite 3 focuses on the new features that were introduced and how it is used both interactively online and computationally.

Chapter 2

Background

Many decades of research in molecular biology resulted in the availability of vast amounts of data, including not only genomic sequences, but also protein sequences, structural data, and protein meta-data as, for example, functional domain information and interaction data. Unfortunately, the sole availability of data like these does not mean that we have an understanding of what *all* the data means in a broader context. Originally, only single entities of data (e. g. single molecules) have been detected and analysed. With the advent of new experimental techniques it was possible to enrich these pieces of data with additional information. One of the most important breakthroughs in this context was the rise of experimental techniques that allow the detection of interactions of molecules. The molecular apparatus of a cell is mainly controlled by interacting proteins and manipulation of one another. Detecting and understanding direct interactions is a first step to a broader view of biological systems. Putting multiple interactions in context and viewing them as a network of interacting entities brings us yet a little closer to a full understanding of the system.

This chapter introduces both these two ways of viewing biological systems exemplified by topics fundamental for the understanding of later parts of this work. First of all protein-protein interactions in general are described, followed by an overview of one specific type of interaction that is of special relevance in cell signalling: *Phosphorylation*. Then, the DNA damage response is described as a relevant example of a complex network of interacting proteins. At the end of this introduction a number of computational phosphorylation prediction methods will be described and compared. All these parts serve as a basis for later chapters. Specifically, for Chapter 3 on page 30 where Scansite 3, a motif-based phosphorylation-prediction tool, is described, and for Chapter 4 on page 56 which illustrates a method that is applied to the enrich the DNA damage response protein-protein interaction network with novel and potentially relevant interactors.

2.1 Protein-Protein Interactions

The control of cellular mechanisms is highly dependent on the interaction of molecules within the cell. These include DNA-RNA interactions, interactions of proteins with DNA and RNA,

interactions with metabolites, and, of course, protein-protein interactions. For this work, the interactions of interest are in this latter category. De las Rivas and Fontanillo (2010) define protein-protein interactions as "specific physical contacts between protein pairs that occur by selective molecular docking in a particular biological context". This definition includes three very important statements. Starting from the end of the definition they are:

- **Biological context**: Protein-protein interactions are highly dependent on their molecular surroundings (e. g. cell type, co-factors, binding partners, etc.), protein modifications, and a cell's cell cycle phase. This also means that the sole interaction-wise compatibility of proteins and proof of an interaction *in vitro* is not sufficient to show that they in fact do interact *in vivo*. Both interaction partners have to be at the same location at the same time (colocalisation), they have to be ready for an interaction (activated, in a complex, etc.), and they must be biochemically compatible.

- **Selective molecular docking / Specific physical contact**: With biochemical compatibility comes the option of selective molecular docking. Molecular docking describes the direct physical contact of molecules of the interacting parties. Physical contact includes binding of the proteins and post-translational modifications. Proteins bind each other in order to form functional units. These can be divided into stable / permanent complexes which are bound to each other until the protein is degraded and transient complexes that only stay together for as long as their functional influence is needed.

  Protein modifications occur if one protein manipulates the molecular structure of another protein, most often in the form of proteolytic cleavage (manipulation of peptide bonds) or covalent modifications (e. g. binding of small molecules to a protein). The former events are mostly irreversible, whereas the latter ones often are reversible (Blom et al., 2004). For example, if a phosphate group is covalently bound to a protein, this group can be removed again by another protein. This kind of modification is called phosphorylation and will be described in more detail in Section 2.1.3. Other examples for covalent modifications are ubiquitination (the covalent binding of ubiquitin, a small regulatory protein), glycosylation (the attachment of a carbohydrate), and sumoylation (attachment of a *s*mall *u*biquitin-like *mo*difying (SUMO) protein). Modifications, especially reversible ones, can be viewed as molecular switches that play a central role in cellular signalling pathways.

  It is very important that protein modifications do not occur randomly whenever proteins that are enabled to modify others meet. For this reason the definition includes the specificity of protein-protein interactions. This means that the interaction specific sites on the protein should be non-generic, meaning that they serve a specific purpose and thus only *recognise* specific interaction partners (De las Rivas and Fontanillo, 2010).

- **Between proteins pairs**: Obviously, the definition of protein-protein interactions only includes the interactions of proteins with one another, excluding interactions with other molecule-types like DNA, RNA, metabolites, cofactors, or ligands. Nevertheless, these kinds of interactions are also very important when studying biological systems as they provide the molecular surroundings that are crucial for many PPIs (see above). This definition also does

not fail to mention that protein interactions can always be viewed as binary interactions. Even though interactions are often described as *complex A* interacts with *protein X* this does not mean that every protein involved in *complex A* interacts with the *protein X*: Every interaction can be broken down into a set of pairs of interacting proteins. This definition also excludes interactions of a single protein with itself (e. g. forming of disulphide bridges between two cysteine residues or autophosphorylation).

2.1.1 Experimental Identification of Protein-Protein Interactions

There are many different experimental methods to determine and verify protein-protein interactions. They can be divided into two groups dependent on their throughput. Low-throughput methods are the most accurate and reliable ones; high-throughput methods have the advantage of being fast, inexpensive, and scalable.

The most widely used high-throughput method for detecting protein-protein interactions are Yeast-2-Hybrid (Y2H; originally described in Fields and Song (1989)) and Tandem Affinity Purification Mass Spectrometry (TAP-MS; see Rigaut et al. (1999)). Y2H is based on the idea of separating a transcription factor's DNA-binding and activation domain across the two molecules of interest so that only their interaction activates the expression of a reporter gene: The DNA-binding domain of the transcription factor is bound to one of the potentially interacting molecules (the so-called *bait*) and the activation domain is bound to the other one (the *prey*). An interaction of the molecules activates the transcription factor and a reporter gene is expressed. As this method is applied in yeast *in vivo*, false positives can occur: It is possible that the binding between the two proteins of interest does not happen directly but via an intermediate yeast protein. Another caveat of this method is that although it has the potential of proving real *in vivo* interactions, the interactions that are proven may never happen naturally in a living system as the two molecules may never meet (Koh et al., 2011). The *bait* and *prey* naming convention can also be applied to TAP-MS. Bait molecules are captured in a matrix that a mixture of prey molecules is passed through. Only those molecules with high affinity to bait molecules (i. e. interactors) will stay in the matrix. This step is called (tandem) affinity purification. The actual identification of interactions is done by mass spectrometry. But there are also drawbacks to this method: Most importantly, this method reports primarily stable and high-affinity binding interactions since weak bindings and transient interactions are probably lost in the matrix-washing process (Pflieger et al., 2010). In addition, TAP-MS may also report interactions that are not relevant in the context of a living system.

An experimental method that is accepted as quasi-gold standard for reporting protein-protein interactions is X-ray crystallography. This method is extremely low-throughput and requires a lot of expertise to be carried out. It is limited to water-soluble proteins and requires a large quantity of extremely pure samples. However, it has the potential of giving very detailed information (atomic level) about the site of interaction and type of bonds (Koh et al., 2011).

Although all commonly used protein-protein interaction detection methods have caveats that have to be considered one can argue that the confidence of a reported interaction is higher if different

experiments have reported this interaction to be *real*. A problem with this rule of thumb is however that negative results are rarely reported in the literature and thus non-verified interactions due to failed verification experiments are not available in public repositories. This also means that experiments showing opposing results about previously verified interactions are usually not made public. In contrast, many public databases exist that collect experimentally verified protein-protein interactions.

2.1.2 Protein-Protein Interaction Databases

The emergence of high-throughput techniques for identifying protein-protein interactions naturally meant a rise in the number of experimentally verified and published interactions. With this came the need for publicly accessible collections of reported interactions. In the past ten to fifteen years a number of protein-protein interaction databases have become available. Based on their content, they can generally be divided into two categories: Those that only contain experimentally validated interactions (*in vivo* or *in vitro*) and those that include computationally predicted interactions (*in silico*). In this section the focus lies on the former category. A selection of major data-repositories is shown in Table 2.1. With the exception of the STRING database all of the databases listed here are curated databases, i. e. the interactions stored there were manually extracted from the literature by a team of biologists or submitted by the publishing group. The curation process is very slow and inefficient; however, it is the least error-prone attempt of tackling the problem of filling such data repositories. Only controlled submission-strategies can guarantee that only valid data is stored. Literature-mining algorithms (e. g. the Rule-based LIterature Mining System for protein Phosphorylation (RLIMS-P) introduced in Hu et al. (2005)) are very efficient but are, generally speaking, error prone and include many false positives and miss true positives. In addition to data from curated databases (including the ones listed in the table), the STRING database also includes data found by literature mining algorithms.

Table 2.1: This table shows an overview of the major protein-protein interaction currently available. Numbers as of April 19th 2012. Dashes (-) denote that the databases do not offer a number in this category. Please note that the STRING database is not a primary protein-protein interaction database itself, but stores the interactions from a collection of different PPI databases: Among others, STRING includes all the PPI-databases listed here.

Protein Protein Interaction Databases

Name (Reference), Link	Proteins	Interactions	Publications
BioGRID (Stark et al., 2011), thebiogrid.org	42,195	334,644	30,884
DIP (Salwinski et al., 2004), dip.doe-mbi.ucla.edu	24,430	73,268	-
HPRD (Prasad et al., 2009), hprd.org	30,047	39,194	453,521
IntAct (Kerrien et al., 2012), www.ebi.ac.uk/intact	61,542	292,919	5,366
MINT (Licata et al., 2012), mint.bio.uniroma2.it	35,048	240,760	130,744
STRING (Szklarczyk et al., 2011), string-db.org	-	-	-

A very important feature of all databases of this type is a link to the publications (i. e. references) and

experiments where the interaction information was published. The Biological General Repository for Interaction Datasets (BioGRID), the Database of Interacting Proteins (DIP), the Human Protein Reference Database (HPRD), the InterAction database (IntAct), and the Molecular INTeraction database (MINT) all contain this information. STRING (Search Tool for the Retrieval of Interacting Genes) saves both the initial origin of the interaction information and the database. All of these resources also include some useful tools for analysing, viewing and downloading interaction data.

Given the fairly high numbers in Table 2.1 for the numbers of proteins and interactions covered by the curated publications it is a surprising fact that there is not a large inter-database overlap in the reported interactions. Turinsky et al. (2010) reported that from about 15,500 publications that are shared across nine major public protein-protein interaction databases only 42% of the interactions and 62% of the proteins curated from the same publication fully agree on average across two databases. The authors conclude that this is mainly due to divergent curation policies across databases, but that protein-identifier mapping (e. g. mapping of isoforms, mapping from protein IDs to gene symbols) and ambiguous phrasing in scientific texts also play a role. A similar result was shown in De las Rivas and Fontanillo (2010) where the authors report that in their comparison of six major protein-protein interaction databases only three interactions were shared across all of these databases. These studies suggest that a combination of databases is the best choice when the best possible set of interactions is needed. This is where the main advantages of the STRING databases come into play: (1) It combines the information from many interaction databases and (2) it provides a score for each interaction based on the number and quality of the reported interactions between the two potentially interacting partners.

Several strategies for solving the issue of divergent methods of manual curation have been suggested. The most promising of which is the idea of coupling the interaction-saving process to the process of paper submittal. If all scientific journals require PPIs to be reported in a standardised way for publishing, ambiguities and other related problems could easily be avoided. This process is already implemented for publishing nucleotide sequences (Mathivanan et al., 2006). Other ideas to unify and standardise protein-protein interaction information have been published and are used (Hermjakob, 2006) but have not changed much in the current body of databases.

2.1.3 Protein Phosphorylation

One of the most abundantly happening covalently modifying protein-protein interactions is *protein phosphorylation*. This term describes the biochemical activity of transferring a phosphate group (PO_4) from an adenosine triphosphate (ATP) to a hydroxyl group (OH) at an acceptor residue turning the ATP into ADP (adenosine diphosphate). Acceptor residues are all amino acids that contain OH-groups, i. e. serines (S, Ser), threonines (T, Thr), and tyrosines (Y, Tyr) at the phosphorylated protein, the *substrate*. The process of phosphorylation is catalysed by enzymes called *protein kinases* and usually occurs in the nucleus or the cytosol of a cell. The reverse operation (removing a phosphate group from an S, T, or Y residue) is carried out by *phosphatases* and referred to as *dephosphorylation*.

Researchers started paying attention to the role of phosphate metabolism when they noticed the huge turnover of phosphate in a wide variety of living cells in different kinds of tissue. The presence of phosphoserines and phosphothreonines (i. e. phosphorylated serine and threonine residues, respectively) suggested that there was some enzymatic mechanism that is responsible for the *use* of phosphorus in the sense of phosphorylation and dephosphorylation. However, almost nothing was known about the proteins that carried out these modifications. Based on these basic ideas and assumptions, Burnett and Kennedy (1954) carried out experiments that reported the "finding of an enzyme that was capable of [...] catalyzing the phosphorylation of a protein substrate by ATP". Simply put, they marked ATP radioactively with P^{32} and found the radioactive phosphorus isotopes again at phosphoserines of the substrate. This finding showed the first evidence of a protein kinase and started a large number of studies on phosphorylation.

In the meanwhile, more than 500 human kinases are known and many of them have been studied extensively, whereas the function of others is still to be determined (Hutti et al., 2004). They are involved in all kinds of cellular processes, including the regulation of binding affinities, the alteration of gene expression by transcription control, manipulation of protein activities in signal transduction pathways, and cell cycle regulation. Many kinases have been originally found in one context, but further in-depth studies showed that they seem to be primarily involved in other important regulatory functions. A good example for a kinase like this is the ATM kinase. It was originally found in context of ataxia-telangiectasia (A-T) patients. Researchers found that A-T patients had a mutation of a gene in common that encodes a kinase, hence naming it ataxia-telangiectasia mutated or ATM kinase (Lee and Paull, 2007). It was not until later that researchers discovered one of the main roles of ATM, which is the regulation of the DNA damage response after double-strand breaks. This particular example will be discussed in more detail in Section 2.2.

The definition of protein-protein interactions described in one of the preceding sections stated that the specificity of the interacting partners plays a central role. Of course, this also applies to kinases. It seems obvious that different kinases prefer to phosphorylate different substrates. But how do they distinguish between different substrates? Clearly, the kinase-substrate interaction focuses on the acceptor residue in the substrate sequence. Crystallisation studies showed that the amino acid sequence around the potentially phosphorylated site plays an important role for the specificity of kinases *in vivo*. It was shown that around nine to twelve residues on the substrate are likely to physically contact the kinase's active site (Songyang et al., 1994) suggesting that approximately this part of the substrate's primary structure determines whether an acceptor residue is likely to be phosphorylated by a given kinase or not. The sequence that is crucial for the biochemical decision of the kinase to do so (or not) is thus dependent on a kinase-specific *consensus motif* / *consensus sequence*. This term describes the amino-acid preferences of kinases around an acceptor residue and includes the kinase's preference for one or more acceptor residues. For example, kinases that prefer serines and threonines as acceptor residues are called serine-/threonine-kinases (e. g. ATM kinase); kinases that only phosphorylate tyrosine residues are referred to as tyrosine-kinases (e. g. the epidermal growth factor receptor (EGFR) kinase).

The amino acid specific preferences of kinases around the acceptor residue can be described by a position specific scoring matrix (PSSM). A PSSM like this contains a probability value for each

amino acid at each position around the acceptor residue, dependent on the kinase's preferences. If, for example, a kinase only phosphorylates tyrosines that are followed by a proline residue (YP), the PSSM will contain a value of 1 for tyrosine and proline at positions 0 and $+1$, respectively (assuming that the tyrosine is at position 0 in the matrix). If there are no preferences at all for the rest of the positions around the acceptor tyrosine all other position's values in this hypothetical matrix would be preferred equiprobable and thus contain a value of $\frac{1}{20}$ (considering a number of 20 amino acids in this matrix). Table 2.2 shows an example for a PSSM like this that additionally prefers a negative residue in position -3 relative to the acceptor site.

Table 2.2: An example for a position specific scoring matrix that describes a kinase's consensus motif three residues around $(+/-)$ the substrate's acceptor residue. The vertical axis describes the positions around the acceptor site, the horizontal one the amino acids (leaving out H, I, K, and L). The motif described here shows a requirement of tyrosine as an acceptor, a requirement of proline at $+1$, and a preference for a negative residue at position -3, favouring aspartic acid (0.6) over glutamic acid (0.4). There are no preferences for other positions (all 20 amino acids are equiprobable, hence values of $\frac{1}{20}$). The motif shown in this matrix can be written in a simplified form as *[DE]-X-X-Y-P-X-X* , with *X* being *any* residue.

A Sample PSSM

	A	**C**	**D**	**E**	**F**	**G**	...	**M**	**N**	**P**	**Q**	**R**	**S**	**T**	**V**	**W**	**Y**
-3	0	0	0.6	0.4	0	0	...	0	0	0	0	0	0	0	0	0	0
-2	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	...	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$
-1	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	...	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$
0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	1
+1	0	0	0	0	0	0	...	0	0	1	0	0	0	0	0	0	0
+2	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	...	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$
+3	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	...	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$

A recent publication that presented the consensus motifs of five major mitotic kinases (CDK1 / cyclin B, Aurora A and B, Nek2, and PLK1) showed that kinases not only have motifs that they prefer, but that they also have anti-motifs that they disfavour (Alexander et al., 2011). The authors show that the kinases they studied exist in "two functionally orthogonal spaces": The localisation space and the motif space. The former refers to cellular compartments a kinase can be found in, the latter to the consensus motif a kinase favours. In the context of this study, each major mitotic kinase overlaps with every other kinase in at most one of these two spaces. Alexander et al. reason that this is because of the unusual kinase localisations during mitosis. In contrast to other cellular process where the subcellular localisation of proteins is more or less constant, the localisation of proteins during mitosis gets completely mixed up, when the genomic content of the nucleus is partitioned into two daughter cells. The existence of anti-motifs shown in this study suggests not only another layer of kinase specificity but also a cellular security system that makes sure that kinases do not get in each other's way.

One of the most commonly used experimental techniques for determining a kinase's consensus sequence is OPLS: In Oriented Peptide Library Screens the kinase of interest is mixed with

about 2.5 billion distinct peptides. All these peptides are fifteen residues long and contain a single phosphorylatable residue at their centre position. These peptides are then sequenced and the preference values for each amino acid at each position relative to the phosphorylated acceptor site can be calculated (Songyang et al., 1994). A related method, which was developed about ten years later, improved the efficiency of this approach. It uses a set of 198 distinct peptide mixtures with each mixture having one of the twenty naturally occurring amino acids fixed at each of nine positions surrounding an also fixed central acceptor residue and with the remaining positions being degenerate. Radioactively labelled ATP is then added to the mixtures with the kinase of interest. After this step the mixture is washed to remove the non-phosphorylated peptides and the extent of phosphorylation at each of the positions is determined by measuring the content of radiolabeled phosphate. This method is referred to as Positional Scanning Oriented Peptide Libary Screen or PS-OPLS (Hutti et al., 2004).

The abundance of experimentally verified phosphorylation sites and the importance of phosphorylation as a cellular process lead to the emergence of public data repositories dedicated to phosphorylation sites alone. An incomplete list of databases in this category is shown in Table 2.3. Phospho.ELM contains exclusively phosphorylation sites from scientific publications and phosphoproteomic analyses. Both, PhosphoSitePlus (PSP) and PHOsphorylation SIte DAtabase (PHOSIDA) started originally as databases for phosphorylation sites only (hence the names), but also store other post-translational modifications in the meanwhile, including glycosylated, acetylated and differently modified residues. All modifications in PHOSIDA come from high resolution mass spectrometry experiments carried out by the group that hosts the database. The most comprehensive of these resources, PhosphoSitePlus, contains *in vivo* sites that were manually curated from the literature, but also *in vitro* sites. All of these databases cover more than one organism.

Table 2.3: This table shows an overview of the major phosphorylation-site databases currently available. Numbers as of April 19th 2012. There is no number of publications available for PHOSIDA as this resource exclusively stores post-translational modifications reported by mass spectrometry experiments carried out by the group hosting the database. The databases PHOSIDA, Phospho.ELM, and PSP are available online at phosida.com, phospho.elm.eu.org, and www.phosphosite.org, respectively.

Phosphorylation Site Databases

Name (Reference)	Proteins	Phosph. Sites	Publications
PHOSIDA (Gnad et al., 2007)	23,769	70,095	-
Phospho.ELM (Dinkel et al., 2011)	8,698	42,914	3,657
PhosphoSitePlus (Hornbeck et al., 2012)	18,887	170,213	13,820

The availability of resources like these allows researchers all over the world to access this information and incorporate them in their studies. Although the importance of phosphorylation in cellular functions has been emphasised many times so far, no relevant examples have been described in more detail. This will change in the next section where the DNA damage response is described.

2.2 The DNA Damage Response

So far, only binary protein-protein interactions have been discussed. By zooming out a bit from this very detailed view of cellular signalling and taking a look at a series of PPIs one after another it is possible to assign a meaning to interactions in a broader context. Such a series of protein-protein interactions can be viewed as a signalling cascade (or signalling pathway). Starting with the assumption that cellular signalling cascades are strictly linear, a trigger event at one side of a cascade leads to some outcome at the other end. Given the number of intermediate steps between the initial trigger of the pathway and the result one might ask why cells use a complicated way like this for controlling basic functions. One of the main reasons for this is that signalling cascades allow a very fine-grained regulation of processes. In contrast to direct on/off-switch-like relationships, a cascade of events allows much more points to control the ultimate outcome. This view of biological systems is very helpful for us to get a better understanding of what is going on, since it reduces the system's complexity. However, in reality pathways very rarely are linear one-after-another cascades. Usually, a more appropriate view is that of a network of interactions. In this context, signalling pathways can be considered as specialised parts of an interaction network, and the network consists of several pathways that interact or overlap with one another. In this section the DNA damage response network will exemplify many of the ideas and concepts that have been introduced so far.

DNA (deoxyribonucleic acid) is probably the most important molecule in living organisms as it stores the organism's genetic blueprint. In order to be able to maintain the stability of its genome, it is crucial for a cell that its copy of the DNA molecule stays as constant as possible over the course of many cell cycle iterations. Damages to this complicated macromolecule can have severe consequences (as, for example, different forms of cancer) and, unlike other molecules in the cell, damaged DNA cannot be simply replaced by a new instance. Hence, it is critical for the organism to survive that damages to the DNA are repaired or that cells with damaged DNA are eliminated. This is achieved by a large number of proteins that are able to sense damage (*sensors*), mediate signals (*mediators*), and repair damage (*effectors*). Each of these groups of proteins contains a number of members. Which ones are activated and which ones they activate in further consequence is highly dependent on the type of DNA damage and how this damage was induced. For example, one of the most dangerous types of DNA damage is when both strands of the DNA break. This type of damage is called *double strand break* and, dependent on what caused this kind of damage (e. g. ultraviolet vs. ionising radiation), a different repair pathway (including different proteins) is pursued.

2.2.1 Types of DNA Damage and Repair Strategies

Threats to the integrity of DNA can be divided into three groups. First of all, environmental (or exogenous) chemical or physical agents pose a threat, including exposure to ionising radiation (IR) and ultraviolet light (UV) on the physical side and cisplatin, mitomycin C (MMC) and other agents on the chemical side (Ciccia and Elledge, 2010). Secondly, products of normal or dysregulated cellular metabolism may interact with DNA and change it biochemically. This includes mainly

highly reactive oxygen and nitrogen metabolites, but also alkylating agents and other metabolites. Third, damage may happen any time due to replication errors or spontaneous reactions, especially hydrolysis (Hoeijmakers, 2009).

All these threats can lead to different types of DNA damage and initiate a number of different repair pathways. All of these pathways are part of the DNA damage response network, i. e. the network of interacting proteins involved in the DNA damage response. Damage to the DNA can be split into three groups: DNA strand breaks, nucleotide-related damages (mostly base adducts), and interstrand crosslinks.

DNA strand breaks are either double strand breaks (DSB) or single strand breaks (SSB). Single strand breaks are breaks in just one of the two strands of a DNA helix. They are repaired by a pathway referred to as single strand break repair (SSBR). Since only one of the two strands is damaged, SSBR is carried by nucleases which use the intact strand as a template. Since this option is not available for DSBs, the repair of double strand breaks is more complicated. At the moment of writing this, two different DSB repair pathways are known: homologous recombination (HR) and non-homologous end joining (NHEJ) (Weterings and Chen, 2008). HR is a very accurate process. It uses a homologous sequence as a template for repairing the break. In contrast, NHEJ ligates the ends of the broken DNA molecule together without using a template. This often results in insertions, deletions or base pair substitutions (Ward and Chen, 2004). Figure 2.1 illustrates the mechanisms of homologous recombination and non-homologous end joining.

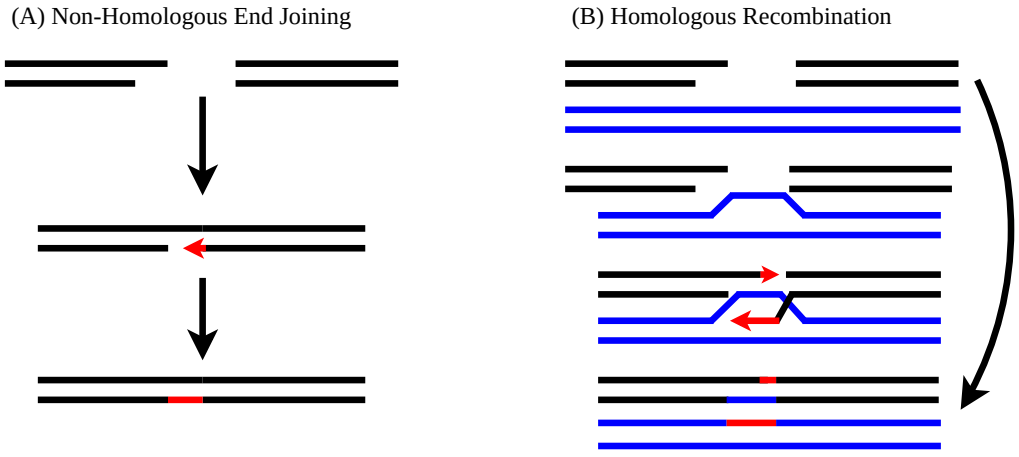


Figure 2.1: Simplified overview of non-homologous end joining (NHEJ) and homologous recombination (HR). The broken DNA is coloured black, newly synthesised strands are coloured red. (A) NHEJ ligates the two loose pairs of strands together. If the gap in one sequence is longer than in the other, the longer sequence is used as a template to complete the other strand. This might result in errors, as it is not guaranteed that the DNA sequence was not altered. There may be nucleotides missing, wrong, or added in the repaired version. (B) HR is much safer and more accurate. It uses a sequence from a homologous sister chromatid as a template (illustrated in blue). This way, there is a template for both strands available and it is possible to correct for deletions, additions or other mutations. In this example, NHEJ is missing a part in the top strand that HR is able to fix.

Damages related to single or multiple nucleotides in the DNA and their repair pathways can also be

divided into different groups: Mispaired DNA bases, helix-distorting damages (e. g. pyrimidine dimers), and damages to single bases (e. g. chemical alterations). Simple damages, like mispaired DNA bases or chemical damages to single bases are repaired by mismatch repair or base excision repair which replace the incorrect or damaged base, respectively. Helix-distorting damages are more complicated lesions that include covalent bindings which obstruct the DNA's helix structure. Damages in this category are repaired by nucleotide excision repair which removes approximately thirty base pairs at the damaged sites and replaces them with new ones (Ciccia and Elledge, 2010).

Interstrand crosslinks are repaired by a pathway called interstrand crosslink repair (ICL-R). In this pathway, the crosslinked DNA strands are first separated and then ligated in the correct way. This process involves a combination of several of the pathways described above (Hoeijmakers, 2009). Figure 2.2 shows a schematic overview of the concepts described so far.

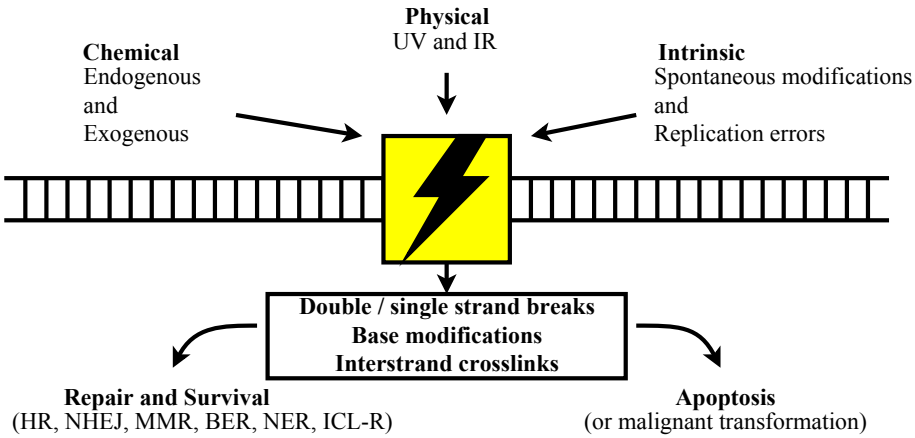


Figure 2.2: A schematic overview of DNA damage agents (top), and a cell's response to them.

2.2.2 The DNA Damage Response in the Cell Cycle

DNA damage is a very common event in the normal lifecycle of a cell. It was estimated that per day and per cell up to 10^5 DNA lesions happen spontaneously (Lindahl, 1993). For this reason it is very important for cells to check the integrity of their DNA on a regular basis. Similar to computers that check the status of their hardware each time they are powered on with the so-called *power on self-test*, cells also have fixed points in time when they perform their DNA integrity checks. These time-points have been found to be three checkpoints during the cell cycle: One at the G1/S transition (G1/S checkpoint), one during the S-phase (intra-S-checkpoint), and another one at the G2/M transition (G2/M checkpoint). The cells fate is decided at these DNA damage checkpoints. This decision process can have three different outcomes: (1) The DNA is found to be in a proper state which allows the cell to continue the cell cycle and the cell goes into senescence. (2) Damages were found and they can be repaired which results in a temporary cell cycle arrest and the activation of the appropriate repair pathway. Or (3) damage was detected, but it is either irreparable or the repair failed. This results in the activation of the controlled cell death pathway (apoptosis). If, for

any reason, this interaction networks fails, mutations and chromosomal aberrations may arise which can cause severe clinical phenotypes.

As for many other signal transduction pathways, the proteins involved in the DNA damage response pathways can be divided into groups dependent on their role: sensors, mediators (or signal transducers), and effectors (Harper and Elledge, 2007). A sensor's job is the recognition of damage and signalling the presence of damage to mediators. These mediators then transmit this information to effector proteins, which then decide the cell's fate (senescence, repair, or death). To give a coarse-grained overview of how a response like this looks like and what genes are involved, the response to DNA double and single strand breaks is outlined in the following.

Sensors. Only minutes after the induction of DNA double strand breaks the damage is sensed by the MRN (*MRE11-RAD50-NBS1*) complex (Reinhardt and Yaffe, 2009). The binding of this complex to the site of damage initiates a cascade of other events. To start with, *ATM* is recruited to the site of damage. This protein is a kinase that plays a central role in the DNA damage response pathway. Mutations in *ATM* (ataxia telangiectasia mutated) cause ataxia-telangiectasia (A-T), a rare but severe neurodegenerative disease, which can be traced back to the deficiency of the cell to react to damages to the DNA. The *ATM* kinase phosphorylates many proteins, including the MRN complex, the histone *H2AFX*, the checkpoint kinase *CHEK2*, the tumour protein 53 (*TP53*), the breast cancer type 1 susceptibility protein (*BRCA1*), and itself (Shiloh, 2003). Another well-studied DNA damage response pathway is centred on the *ATR* (A-T and Rad3 related) kinase and the checkpoint kinase *CHEK1*. This pathway is initiated by the 911 (*Rad9-Rad1-Hus1*) complex after DNA single strand breaks (Cimprich and Cortez, 2008). After recruitment to the site of damage, *TopBP1* binds to the 911 complex. *TopBP1* contains a domain that can bind *ATRIP* and stimulate *ATR* activity. *ATR* binds to *ATRIP* (*ATR* interacting protein).

Mediators. Proteins that are direct substrates (i. e. phosphorylation targets) or temporary binding partners of *ATM* and *ATR* are referred to as mediators in the DNA damage response (Harper and Elledge, 2007). The most important proteins in this category are the checkpoint kinases *CHEK1* and *CHEK2*, *BRCA1*, *H2AFX*, and *MDC1* (mediator of DNA damage checkpoint protein 1). Phosphorylation of the histone *H2AFX* by *ATM* and *ATR* promotes the loosening of the chromatin in that region. *ATM* and *ATR* activate the effector kinases *CHEK2* and *CHEK1* by phosphorylating them, respectively.

Effectors. Downstream of mediators, effectors act on proteins directly involved in cell cycle transitions and repair strategies. These components are what give the checkpoints their unique identities (Sancar et al., 2004). Proteins in this group include the three phosphotyrosine phosphatases *CDC25A*, *CDC25B*, and *CDC25C* that dephosphorylate cyclin-dependent kinases like, for example, *CDK2* as well as *TP53*. These phosphatases are inactivated by phosphorylation which promotes the continuation of the cell cycle. In the G1/S checkpoint cell cycle arrest can be initiated by the phosphorylation of *TP53* by *ATM* or *ATR*. Phosphorylation of *TP53* then activates the transcription of *P21* (or *CDKN1A*), a CDK inhibitor. This, in further consequence, blocks the transcription of genes required for the initiation of the S-phase.

It is important to mention that assigning proteins into these categories does *not* mean that they act only at this stage in the cell cycle and in this function in the DNA damage response. Different cell cycle checkpoints favour different groups of proteins. They differ most in which effectors they use, but some sensors and mediators also play more important roles in some checkpoints than in others. However, many key players are active during the whole response process. ATM, for example, is only one of many kinases that phosphorylates proteins in the early and later response to double strand breaks, independent of which cell cycle checkpoint is currently happening (Sancar et al., 2004). In a similar manner, groups of proteins are often interchangeable, meaning that one group of proteins serves as a backup for another group. Double strand breaks, for example, can either be repaired by a pathway initiated by *BRCA*, or by *PARP* proteins (Hoeijmakers, 2009). The presence of backup repair-pathways plays an important role in some therapeutic strategies as the following section will show.

2.2.3 Relevance of the DNA Damage Response in Diseases

The cellular signalling network outlined in the previous section shows only a strongly simplified and thus incomplete picture of the DNA damage response. It is known that this particular network alone incorporates hundreds of other proteins and molecules. However, many diseases are known that originate from the misregulation or dysfunction of a single entity. For example, A-T is a genetic disorder in humans that occurs if the ATM gene is either lost or inactivated. It belongs to a group of diseases referred to as *genetic instability syndromes* which have in common that they result from a defective DNA lesion response mechanism. Other well-studied diseases in this category are (amongst others) Xeroderma pigmentosum (XP), the Nijmegen breakage syndrome (NBS), and Fanconi's anemia (FA). All of these diseases can cause different types of cancer as a broken damage check and response apparatus can cause uncontrolled cell proliferation and / or avoid apoptosis. Familial breast cancer also results from mutated DDR genes. Table 2.4 sums up the main facts about these diseases noting the mutated genes, what part of the DNA damage response is defect, and what cancer(s) the mutations are known to cause.

Based on the protein interactions described in the previous section it is quite comprehensible that mutations in key players like *ATM* and *NBS1* (part of the MRN-complex) disrupt cells' damage response pathways and thus make them misbehave. In contrast to familial breast cancer however, the primary phenotype of the other diseases mentioned before is not cancer, but neurological (e. g. microcephaly, retardation, ataxia), immunodeficiency-related, some kind of dysmorphism, and / or tissue defects. Xeroderma pigmentosum patients, for example, are extremely sensitive to the UV-light component of sunlight. They show accelerated signs of ageing of their skin and are highly susceptible to skin cancers (Shiloh, 2003). As mentioned before, UV-radiation causes different types of DNA damage. In XP some of the genes that are needed for fixing this kind of damage are mutated (inactivated or lost), which explains the abnormal behaviour of those cells that are exposed to sunlight: the skin cells. One of the main characteristics of FA is bone marrow depletion, which causes insufficient formation of all blood cell types. Cells of Fanconi's anemia patients are sensitive to DNA interstrand crosslinking and breaking agents. It is assumed that all FANC-genes play a role

Table 2.4: Overview of a small number of genetic diseases that were associated with DNA damage response genes and are known to be able to cause cancer. A more complete list can be found in Ciccia and Elledge (2010). The diseases listed here are Ataxia telangiectasia (A-T), Fanconi's anemia (FA), Nijmegen breakage syndrome (NBS), Xeroderma pigmentosum (XP), and familial breast cancer (FBC). The abbreviations used for DDR defects HR, DSB, ICL, and NER stand for homologous recombination, double strand break, interstrand crosslink, and nucleotide excision repair, respectively.

Genetic Diseases Associated with DNA Damage Response Defects

Disease	Mutated Genes	Main DDR Defects	Possible Cancers
A-T	*ATM*	damage signalling, DSB repair	lymphomas, leukaemia, breast cancer
FA	*FANC*-genes	ICL-repair, HR	AML, myelodysplasia, carcinoma
NBS	*NBS1*	damage signalling, DSB repair, replication fork repair	B cell lymphoma
XP	*XPA-G*, *POLH*	NER	carcinomas, melanoma
FBC	*BRCA1*, *BRCA2*, *CHEK2*, *NBS1*, *ATM*, and others	HR, damage signalling	breast cancer, ovarian cancer

in the so-called FA-pathway which includes parts of the HR- and ICL-repair pathways (D'Andrea, 2010).

Increasing knowledge about the protein-protein interactions involved in the DNA damage response network and the extraction of functional modules allows us to target diseases directly at their source. Given that we know what mutations a disease is caused by — and we do know that for many diseases — it is possible to force cells to use alternatives to the malfunctioning gene. One approach that exemplifies this idea quite well is the use of a pathway inhibitor for targeted cancer treatments. In an illustrative scenario where cancerous cells are proliferating in an uncontrolled manner, a specific pathway that enables a cell to repair double strand break damage may be inhibited due to missing or blocked genes that are needed for this pathway's (de-)activation. In healthy cells in the same organism, however, all pathways work fine. Knowing that the cell knows an alternative pathway that serves the same purpose as the one that is damaged in the cancerous cells (i. e. repairing DSB damage) allows clinicians to turn off the alternative pathway by using some specific pathway inhibitor. Consequently, both repair strategies are turned off in the cancerous cells, but one pathway is still working in the healthy cells. Thus, the misbehaving cells are killed, and the healthy ones stay alive. Figure 2.3 illustrates this approach.

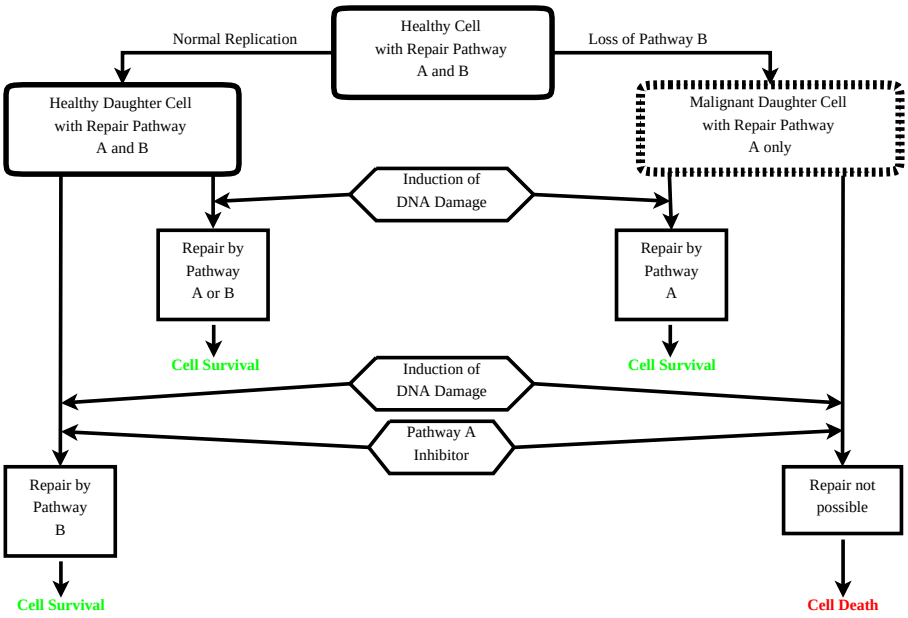


Figure 2.3: Illustration of the basic idea of pathway inhibitor based therapies. Here, healthy cells have two DNA damage response pathways: A and B. If one pathway is lost for some reason (in this case B), cells may become malignant cancer cells (top right). These cells can be forced to apoptosis by inhibiting their backup repair pathway (A). This will kill cancerous cells (bottom right), but healthy cells will stay alive (bottom left).

2.3 Computational Methods for Predicting Kinase-Substrate Interactions

The biological importance of protein phosphorylation was emphasised many times so far and techniques that allow researchers to experimentally verify protein-protein interactions were outlined. But even high-throughput techniques lack the ability to answer some scientific questions in a resource- and time-efficient manner. For example, given a species' entire proteome, what are a kinase's potential targets? Or, which kinases are likely to phosphorylate a set of given proteins? Although it is indeed possible to answer these questions experimentally, a more resource-responsible way is to use computational tools that use experimental data to predict kinase-substrate interactions. Almost 40 different phosphorylation prediction tools (counting only phosphorylation-specific tools, excluding more general PTM-predictors) have been published since 1999 (Trost and Kusalik, 2011). This section introduces some representative tools that are available online focusing on their prediction method. It is obvious that all of these methods are based on prior biological knowledge and experimental data. The nature of the underlying data varies, though. Some methods focus on a specific biological feature (e. g. a kinase's consensus motif); others incorporate a wider variety of features (e. g. different kinds of sequence properties, protein meta-data and chemical properties), again others utilise phosphorylation-relevant environmental data (e. g. known protein-protein interactions), and some methods combine all these kinds of information.

Scansite[1] (Yaffe et al., 2001) is one of the oldest and most widely used kinase-substrate prediction tools. It uses position specific scoring matrices (PSSMs) that describe different kinases' recognition motifs to identify potential phosphorylation sites in proteins. In addition, users are presented with some supporting information about the queried proteins which assists them in their decision process of whether to trust a prediction. This method will be described in detail in Chapter 3 on page 30. PSSMs are used by most of the phosphorylation site prediction methods in some way. The DISorder-enhanced PHOSphorylation predictor *DisPhos*[2] (Iakoucheva et al., 2004) is based on the fact that many important regulatory proteins contain intrinsically disordered regions and phosphorylation sites in proteins of this category are more likely to occur within intrinsically disordered regions. Intrinsically disordered proteins are proteins that contain intrinsically disordered regions and thus do not have a rigid 3D structure. Hence, proteins like this undergo conformational changes over time. It is estimated that up to 30% of all eukaryotic proteins are mostly intrinsically disordered and that around 70% have long disordered regions (Uversky and Dunker, 2010). In addition to five predictors of disorder DisPhos takes into account sequence information (e. g. secondary structure, surface accessibility, charge, flexibility). This information is collected for a set of training data, followed by a feature extraction step that identifies the most important features. These are then used to train a predictor that is used to predict phosphorylation sites (Iakoucheva et al., 2004). Another tool, ConDens[3] (Lai et al., 2012), focuses on evolutionary conservation and does not depend on training data. This tool is based on the idea that a kinase's recognition motif is not specific to a single species but is evolutionary conserved. Consequently, important phosphorylation sites are evolutionary conserved too. But this idea is not as straight-forward as it seems. It is not enough to check if the potentially phosphorylated site is evolutionary conserved. The whole recognition motif has to be conserved. Several other issues have to be considered too: Is a site conserved because it is important, or because the stretch of the amino acid sequence where the site resides is conserved? Where is the site in another organism after deletions and insertions? How can functionally conserved regions be identified? How to deal with point mutations of Serines, Threonines or Tyrosines to other amino acids? ConDens deals with these evolution-related issues by using an evolutionary model to account for local sequence divergence in the multiple sequence alignment of candidate sequences (Lai et al., 2012). Sites that match a given recognition motif and are evolutionary conserved are more likely to be *real* than others that are not conserved. However, ConDens uses regular-expressions instead of PSSMs to represent motifs, which causes the method to only find very strict patterns and does not allow it to incorporate a kinase's amino acid preferences for specific positions. NetPhosK[4] (Blom et al., 2004) is a tool that uses a pure machine learning approach. It uses artificial neural networks to train a kinase-specific predictor based on known (i. e. experimentally verified) phosphorylation sites. NetworKIN[5] (Linding et al., 2007), the last method that will be mentioned here, utilises data from protein-protein interaction databases (primarily STRING) as well as predictions from other prediction tools, specifically Scansite 2 and NetPhosK. This tool tries to incorporate contextual information in the prediction process. Here, the term

[1] *http://scansite3.mit.edu/*
[2] *http://www.dabi.temple.edu/disphos/*
[3] *http://www.moseslab.csb.utoronto.ca/andyl/*
[4] *http://www.cbs.dtu.dk/services/NetPhosK/*
[5] *http://networkin.info/*

contextual information describes the biochemical surroundings of kinase-substrate pairs as, for instance, subcellular compartmentalisation, colocalisation via anchoring or scaffolding proteins, and the activity of regulatory subunits. In contrast to the tools described so far, this tool cannot be used to predict novel phosphorylation sites, but rather is designed to assign kinases to substrates with experimentally verified phosphorylation sites. This is done by first running the predictor tools Scansite 2 and NetPhosK on a given protein to assign a kinase family to a site and then using contextual information from other databases to determine which kinase is the most likely to be responsible for a site of interest.

When using prediction tools like these it is very important to keep their methods' caveats in mind. To start with, predictions are just predictions and do not guarantee that they are right. This is especially important if tools are used to design and plan experiments. Another caveat cannot directly be blamed on the method that is used, but on the data that the method is using. The data, be it experimentally verified sites or position specific scoring matrices from OPLSs (or other data), originates from experiments and these experiments may be error-prone. This applies especially to methods that use some kind of machine learning approach to train a predictor as they highly dependent on the quality of the underlying training dataset. Obviously, a big set of training data is necessary to create a good predictor and, indeed, a large number of experimentally verified phosphorylation sites is available in public databases. However, training a predictor also requires a negative dataset that gives information of what sites are known to never be phosphorylated. Data like these are usually not published as it is not easy to say if a phosphorylation site was not verifiable due to a failed experiment or because it is just not a phosphorylated site. In any case, negative results are very rarely published in the scientific literature, hence causing a lack of reliable training data for that purpose. The quality and nature of the training data should also be one important thought in the user's mind when deciding whether to trust a predictor that was trained on these data. Amongst others this includes the species of the proteins in the training dataset, the type of experiment that was used to verify the sites, and of course the number of sites and proteins included. Thus, it is very important for prediction tools to (1) give information about how the method works, (2) give a measure that allows users to compare results and distinguish between good and not-as-good results and (3) give additional information that helps the user decide whether to trust the predictions (this information could be incorporated in the prediction method itself, but is also very helpful if it is just presented for the user to examine). Scansite 3, a new version of the previously mentioned Scansite 2[6] does exactly that.

[6] *http://scansite.mit.edu/*

Chapter 3

Scansite 3: A Motif-Based Phosphorylation Prediction Tool

The location of the interface between kinase and substrate is the centre of attention in kinase-substrate prediction tools since they define where phosphorylation sites can be expected. This interface is, however, not solely defined by the amino acid that is going to be phosphorylated, but also by a number of amino acids surrounding this position in the protein sequence. This interface or *recognition motif* is not only the direct physical interface in kinase-substrate interactions but also unique to each kinase and thus the most decisive feature in phosphorylation prediction tools. This is the reason why almost all of the tools described in the previous section take advantage of this feature in some way. One of the most widely used and highly cited kinase-substrate prediction tools, Scansite (Yaffe et al., 2001), focuses on this basic idea and allows users to easily determine if a phosphorylation site can be expected based on a match in the kinase's and substrate's interface, and if so, how likely it is. Scansite was originally developed in 2001 and was one of the first tools in the category of protein-protein interaction predictors. Motif-based tools pre-Scansite searched for plain amino acid patterns in protein sequences: This method, implemented in PROSITE (Bairoch, 1992), was a first step towards matching an amino acid's recognition motif, but was not able to consider different preferences for different amino acids at distinct positions in the pattern, as it was essentially a regular expression search in protein sequences (i. e. a Boolean matching model). Scansite introduced a dynamic matching model that allowed to define a kinase's preferences for the positions in a pattern. In 2003 Scansite was re-engineered (Obenauer et al., 2003). This update introduced some new features, including the option of using user-defined motifs, and allowing users to submit more targeted searches. Scansite 2 also came with an improved user-interface and higher performance. This tool has been used heavily since its introduction: Since usage-recording with Google Analytics[1] was started in 2011, a median of 116 hits a day (on weekdays) was reported. However, the increasing number of kinase motifs that have been experimentally determined over the course of time and the constantly growing protein databases slowly brought Scansite 2 to its limits. These, along with the fact that introducing new features was not easily possible, were the

[1] *http://www.google.com/analytics/*

main reasons why it was decided to give Scansite a new face with the implementation of Scansite 3. All the features that Scansite 2 offered are still available in the new version.

The most important features in Scansite are *Protein Scan* and *Database Search*: Protein Scan searches are scanning a given protein (by sequence or identifier) for sites using either all or some of Scansite's motifs. Database Searches answers the reverse question: Given one (or more) motifs and a protein database, what proteins are likely to be phosphorylated at which location in their sequence? For all of these searches, user-defined motifs can be used. The Database Search feature also offers an option to search for a so-called *Quick Motif*, which allows users to create a simple motif that consists of primary and secondary amino acid preferences at each position on-the-fly. Combined Database Searches with multiple motifs allow different kinds of restrictions, including space restrictions between sites of the different motifs. Other features of Scansite include simple pattern-matching searches in protein databases (*Sequence Match*), a tool for calculating a protein sequence's isoelectric point and molecular weight for different numbers of hypothetical phosphorylation sites (*Calc. MolWeight and pI*), and a tool that calculates and visualises amino acid compositions around different residues (*Calc. Amino Acid Composition*).

This chapter first explains the scoring algorithm used in Scansite, followed by an introduction of the new features introduced in the new version. Later, a number of usage examples explain the features that are available. The chapter closes with a couple of technical details about Scansite 3.

3.1 Motif-Based Scoring Algorithm

Scansite is based on position specific scoring matrices (PSSMs) derived from oriented peptide library screen experiments that represent different kinases' recognition motifs. These scoring matrices define a fifteen amino acid long sequence around the acceptor sites (i. e. seven residues up- and downstream) and define a confidence value for each amino acid at each position. In order to identify potential sites in a given protein Scansite first scans the protein sequence for relevant acceptor residues, i. e. S- and T-sites for serine-/threonine kinases and Y-sites for tyrosine-kinases. A kinase's PSSM is then used to calculate a score for each of these putative sites. First, each position i in this scoring-window w is assigned a position-specific score

$$w_i = \frac{\ln(aa_i)}{\ln(2)} = \ln_2(aa_i)$$

with aa_i representing the PSSM's value that defines the amino acid aa in the window at position i. Given the use of base-two logarithms at that point and the additions in the next step, PSSM-values between zero and one penalise amino-acid occurrences at a given position, values of one show indifference, and values greater than one define favouring affinities. In the next step, these position specific values are summed up into a raw site score s_{raw}, omitting the fixed centre (position 8). Also, this raw score is normalised by the number of scored positions n_{scored} ($0 < n_{scored} \leq 15$). This normalisation is necessary to account for side-effects at acceptor sites that are close to a protein's

N- or C-terminus. The raw score

$$s_{raw} = \frac{1}{n_{scored}} \cdot \sum_{\substack{i=1 \\ i \neq 8}}^{15} w_i$$

is then normalised by a PSSM's optimal score s_{opt}. The optimal score is calculated by summing up the position-wise maxima in a PSSM. In other words, an optimal sequence for a PSSM is the one that, at each position i in the scoring window w, contains the amino acid with the maximum value in the PSSM at this position.

$$s_{opt} = \sum_{\substack{i=1 \\ i \neq 8}}^{15} \max(w_i)$$

This results in the final score s_{final} that is then displayed to the user. The more similar the sequence window that is currently scored is to the optimal motif described by the PSSM, the closer the final score is to zero (numerator). The normalisation by the optimal PSSM-score (denominator) is just a simple method that helps bringing different motifs' scores to a similar range.

$$s_{final} = \frac{s_{opt} - s_{raw}}{s_{opt}}$$

Since every PSSM is different and the scores are highly dependent on the amino acid composition of the scored proteins, Scansite helps users to make sense of the final score by comparing it to precalculated all-proteome scores. More specifically, this means that scores are precalculated and stored for all proteins in a whole proteome and the final site-scores are then compared to the distribution of the proteome-wide scores. The sites that are displayed to the user are filtered by the percentile of best sites in the proteome: The settings high, medium and low stringency are available, which show sites in the proteome's top 0.2%, 1%, or 5%, respectively. In addition, a robust Z-score estimate that is directly calculated from the reference-proteome's distribution is calculated. This value is calculated as

$$z = \frac{s_{final} - median_{referenceProteome}}{MAD_{referenceProteome} \cdot c}$$

with $MAD_{referenceProteome}$ being the median absolute deviation of the reference proteome and c a scaling constant of value 1.4826 which allows to make this measure consistent with Gaussian distributions (Rousseeuw and Croux, 1993) and gives information about how many adapted MADs the site-score is from the reference-proteome's median. In Scansite z is usually < 0 since those sites the top-scoring sites are in the top $\leq 5\%$ percentile. Due to the normalisation of the scores using a motif's optimal score, Scansite's scores start at 0. Consequently this minimum score means that an optimal motif match was found. The higher the score, the more divergence from the optimal motif is found.

It is obvious that this algorithm cannot only be applied to PSSMs that represent kinase-substrate recognition motifs, but also to PSSMs that represent other affinities, e. g. binding specificities. In general, Scansite's motifs focus on S-/T- or Y-residues, but this is only because Scansite was originally designed to predict kinase-substrate interactions. However, Scansite also contains motifs that do not represent a kinase. For example, the Scansite motif *Intersection SH3A* allows to find

potential interaction partners and sites of the SH3 domain protein intersectin 1 (*ITSN1*).

3.2 Improvements in Version 3 of Scansite

Scansite 3 comes with a number of new features and has improved both in terms of usability and maintainability. The centrepiece of Scansite is and always has been the kinase motifs it uses to calculate scores for potential phosphorylation sites. In Scansite 2 a number of 62 motifs are available (Obenauer et al., 2003). Version 3 extends this arsenal with 54 yeast-specific PSSMs and 8 new mammalian motifs. The yeast motifs — as published and described in Mok et al. (2010) — have been provided by the Turk-Lab from Yale University[2]. Since many kinases have been shown to recognise very similar sequences, the 54 PSSMs represent the recognition affinities of more than 60 kinases in yeast (*Saccharomyces Cerevisiae*), i. e. some motifs represent more than just one kinase. For example, *Mck1* (meiotic and centromere regulatory kinase 1), *Mrk1* (Mds1p related kinase 1), and *Rim11* (regulator of IME2 1) most likely phosphorylate a serine or threonine residue that is followed by another serine or threonine residue that is located 4 residues downstream in the substrate's sequence (**[ST]**XXX[ST], with X representing no preference and the bold [ST] being the acceptor residue). Scansite 3 includes all the kinases' names in composite-motifs like these: This particular motif is named *Mck1 and Mrk1 and Rim11*. So far, only mammalian motifs have been available in Scansite. Gene information about the kinases that are represented by these motifs was made available through hyperlinks to the corresponding GeneCard-pages[3] (Stelzer et al., 2011). For yeast motifs in Scansite 3 links to the Saccharomyces Genome Database[4] (SGD) are provided (Cherry et al., 2011). This allows users to quickly get more information about Scansite's motifs.

However, both GeneCard and SGD do not provide any information about how certain motifs look like (other than what acceptor residue some kinases' favour, and even this information is not present for all kinases). In Scansite 3, a new feature allows the visualisation of kinases' recognition motifs in so-called *motif logos*. Figure 3.1a shows the motif logo of the previously mentioned DNA damage kinase *ATM*. Since the logos are derived from Scansite's PSSMs, all motif logos are divided into 15 columns, each of which shows the affinities for those amino acids that the kinase favours in each position. The centre position (pos. 0) represents the acceptor residue. The sizes of the letters do not directly correspond to the values in the PSSSM, but are calculated from these numbers to show the information content that is available all the positions. These numbers are calculated according to the *sequence logo* method published in Schneider and Stephens (1990). In a position that exclusively contains low affinity values, less information content is available than in a position with high(er) affinity values. This directly corresponds to the height of the total of amino acids displayed in one position. Centre positions usually have the highest information content and are, thus the biggest letters. The constantly growing public protein databases also introduced some rarely occurring amino acids that Scansite's motifs needed to deal with: Selenocysteine (U) and Pyrrolysine (O). By default, Scansite 3 treats these amino acids as Cysteine and Lysine,

[2] *http://www.yale.edu/turklab/*
[3] *http://www.genecards.org/*
[4] *http://www.yeastgenome.org/*

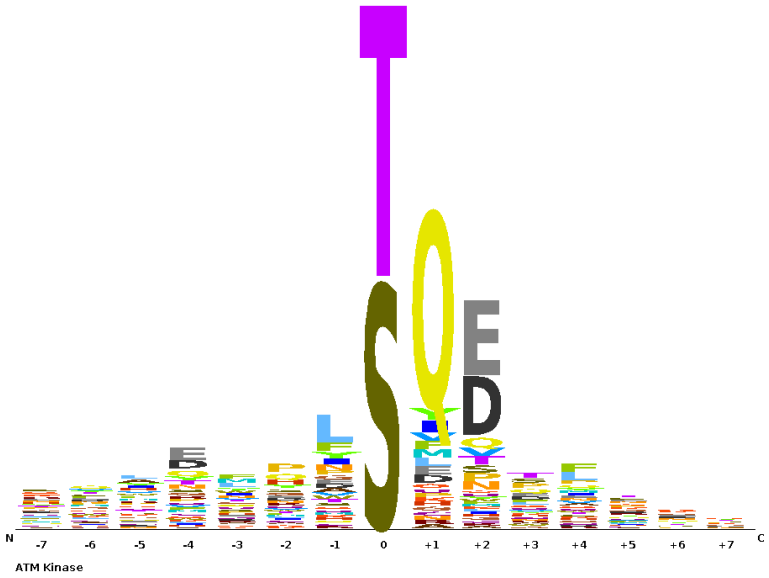
respectively, but motifs are enabled to provide special values for these amino acids. Another caveat when defining recognition affinities is the C- or N-terminus preference that some proteins have, i. e. they only recognise residues at one of the target sequences' ends. This applies, for example, to proteins of the *PDZ domain 1 containing* group (*PDZK1*). For special motifs like this, Scansite 3 displays an exclamation mark (!) for a C-terminus preference and a dollar-sign ($) for an N-terminus preference. *PDZK1*-proteins have been reported to have a preference for a sequence's C-terminus (see Figure 3.1b). The colours that are used for printing different amino acids are chosen to display some of the amino acid's physicochemical properties: Negatively charged amino acids (D and E) are grey, positively charged ones (H, K, O, and R) in shades of red; aliphatic residues (I, L, and V) are displayed in blue-tones, aromatic (F, Y, and W) in green; small amino acids are shown in orange (P and N), purple (C, U and T) and brownish colours (S, A, and G), and the rest either blue-green (M, hydrophobic), or yellow (Q, charged).

In Scansite 2, the quality of predictions is determined by the comparison of site-scores to all scores calculated using a given motif on all putative sites in all vertebrate proteins available in SwissProt[5]. This makes sense for mammalian motifs, but when non-vertebrate motifs are used, it is important to use a different reference-proteome. Hence, alternate reference proteomes are introduced in Scansite 3: Searches for sites with yeast motifs use the yeast-proteome's score distribution (all proteins in the SGD) to calculate a percentile by default, searches with mammalian motifs use the traditional vertebrate-reference. Alternatively, users can decide which reference-proteome to use.

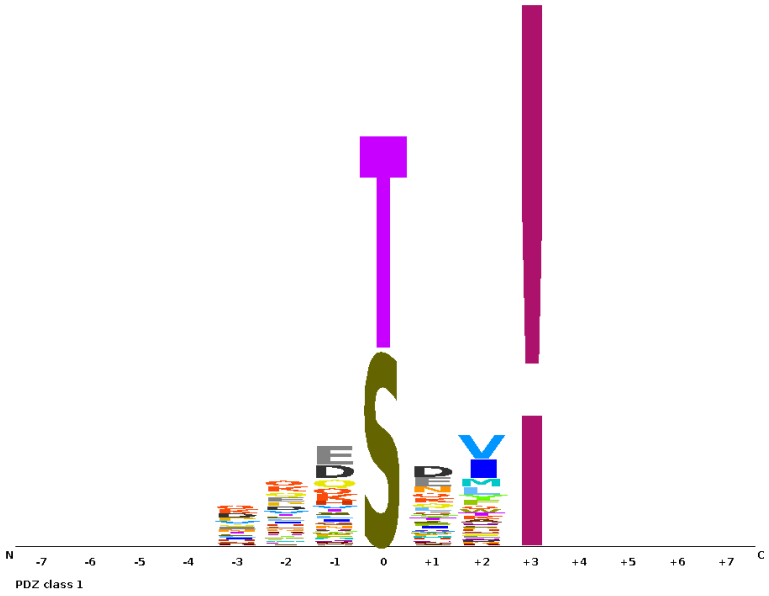
In addition to scores and percentiles for predicted sites, Scansite always presented the user with some additional information about the substrate, including a surface accessibility plot which shows the biochemical accessibility in the folded protein and locations of functional domains as predicted by PFAM[6] (Punta et al., 2012). However, computational access to PFAM often posed a problem in terms of accessibility. This is why it was decided to replace PFAM's predictions with those from InterProScan's PFAM-engine in Scansite 3 (Hunter et al., 2009). This application can be installed locally which circumvents access and latency problems. Although domain starting- and end-positions sometimes differ by a few positions, InterProScan's domain predicitions generally match PFAM's predictions very closely. Positional disagreements may cause confusion, but are not a problem since all numbers come from prediction engines and there is no way to tell which positions are more correct anyway.

One of the most important novelties in Scansite 3 is the introduction of previously mapped sites to support predicted sites. Previously mapped sites are phosphorylation sites from public phosphorylation site databases. At the moment of writing this, the phosphoproteomes of PhosphoSitePlus, Phosida, and PhosphoELM are included (see Table 2.3 and Section 2.1.3 for more information about these databases). If a site predicted by Scansite 3 is found in one or more of these databases, a link to these databases is provided. If the databases support direct links to sites the links take the user directly to the protein- (PhosphoELM) or site-specific (PhosphoSitePlus) content. Otherwise the link refers to the database's basic homepage (Phosida). All previously mapped sites in Scansite 3 are, however, not associated with the kinases that are responsible for those phosphorylation, but

[5] *http://www.uniprot.org/*
[6] *http://pfam.sanger.ac.uk/*



(a) The recognition motif of *ATM*. A strong preference for a **[ST]**Q[ED]-like motif can easily be seen. Please note that the residues D and E in position +2 have similar colours, in this case denoting a preference for negatively charged amino acids in this position.



(b) The recognition motif of *PDZK1*. The exclamation mark at +3 denotes a strong affinity towards a sequence's end.

Figure 3.1: Two examples for motif logos. The larger an amino acid's one-letter-code is printed the higher the preference for this particular amino acid in this position. In many positions the letters are too small to be recognised (especially in ATM's motif logo). Although minor residue-specific preferences like these can usually be ignored as there is no notable information content available from the PSSMs, Scansite's colour codes can help to identify physicochemical preferences.

only serve as supporting information about the reported presence of a phosphorylation of these sites. This means, that this information is not intended to support the kinase-site association, but the confidence of whether a site in general is a phosphorylation site. Nevertheless, it is important to note that a site that is not supported by a previously mapped site can still be a site, and that a site that is supported by data from phosphorylation site databases does not *guarantee* to be true either (but, obviously, increases the likelihood that it is). There are two main reasons for the missing link between previously mapped sites and Scansite's motifs: As with much other biological information, identifiers are often ambiguous, and automated mapping between different types of identifiers is often very complicated and sometimes not possible at all. Secondly, some phosphorylation site databases store data from experiments that report the presence of phosphorylation sites, but do not report the reasons for this phosphorylation. Thus, the kinase(s) responsible for many phosphorylated residues remain to be discovered.

Another important addition in version 3 is the option to access the most important features of Scansite 3 computationally. Scansite 2 was never designed to be accessed computationally since there was no demand for an option like this at the time of publishing it. In the past ten years the use of computational prediction and analysis tools has become a crucial part of every biologists work. Computational access to computational tools allows users to save time by automating parts of their data analysis process. In Scansite 2, computational access was done by submitting HTML-forms and parsing the HTML-response for the relevant information. This technique generally works, but can break programs easily when, for example, the layout (and, thus the HTML-page) of result-displaying pages is changed. In order to make computational usage of Scansite easier, Scansite 3 offers a RESTful (REpresentational State Transfer) web service. This makes accessing Scansite computationally much more programmer-friendly: All parameters that are relevant for a search are defined and submitted in a single URI; the results are returned in XML-format.

Since the public protein databases that Scansite uses as primary data sources have grown remarkably in the past years, some performance enhancement on the server-side was needed to deal with these amounts of data. Most of all, this applies to database searches as their time-wise performance highly depends on the number of proteins returned by the search. Scoring single sites in a protein is an easy task and not at all computationally expensive, but the scoring of thousands of proteins and hundred thousands of sites can take a lot of time and computational power. This is why actions like these were parallelised in an easy manner: For database searches and for creating reference histograms (when a whole proteome is scored), the proteins that are going to be scored are distributed to a number of threads and scored in parallel. The results are then collected again by the initial thread.

In addition to these changes in the web application, a number of tools have been implemented that make maintenance of Scansite 3 much easier. To start with, applications have been created that allow easy population and updating of the data-backend. At the moment of writing this, support for automatic download, parsing and storing of data from SwissProt, TrEMBL[7] (Translation of EMBL nucleotide database), SGD, NCBI Protein[8] (GenPept), and Ensembl[9] (Human and Mouse)

[7] *http://www.uniprot.org/*
[8] *http://www.ncbi.nlm.nih.gov/protein/*
[9] *http://www.ensembl.org/*

is implemented. By using the Java Reflection API it is easily possible to add support for other databases or user-defined protein datasets. Reflection also makes it possible to use alternative domain prediction engines: At the moment, InterProScan can be accessed in two different ways, either on the same computer the web-server is running on, or on a remote computer. In the latter option, a SSH connection is used to run InterProScan on a remote system.

3.3 Features and Use Cases

There are several ways Scansite can be used to predict kinase-substrate interactions. This section will introduce some use cases that exemplify the usage of the web application and the web service, describing not only the user interface interactions that are required to perform searches, but also explaining the results that Scansite 3 returns. In addition, at some points a short overview of what happens on the server-side when submitting a search will be given.

3.3.1 Using the Web Application

The Scansite 3 web application was designed to be as intuitively useable as possible, while at the same time staying close to the general design and feature-naming convention that was used in Scansite 2. When accessing the Scansite 3 homepage[10] a welcome-page will be displayed that offers links to the main search features. Alternatively, these features can be accessed using the buttons in the navigation section on the left-hand side of the page (see Figure 3.2). The link at the top of the navigation section should encourage users to give feedback about their experiences with Scansite 3. This includes bugs they discovered, ideas about how to improve this page, and things they like or dislike about the site. The news section will keep users up-to-date about new features, new motifs, and other changes in Scansite 3. A very important new feature is the section labelled *Databases and Motifs*: Here, users can view motif logos (as described in Section 3.2), motif-group assignments and get an overview of the protein databases available, including their version (or date of last update) and each database's size (i. e. the number of proteins stored). The bottom part of the navigation section allows *administrators* and *collaborators* to login. Collaborators have the privilege of using *non-public* motifs, i. e. motifs that are not accessible to everyone (because, for example, they are based on unpublished data). Administrators have the additional options to add, update, and delete motifs, motif-groups, and news-entries. Another group of administrators, the so-called super-administrators, have the additional privilege of editing user-information (create, update, and delete Scansite-users).

Other links in the navigation section refer to a list of frequently asked questions (*FAQ*s, such as "How to cite?" or "How to create a motif?"), a tutorial (*Tutorial*), and a short overview of what Scansite is and who was involved in creating it (*About*). The *Scansite Features*-links provide access to all of Scansite's features which will be described in detail in the following sections.

[10] *http://scansite3.mit.edu/*

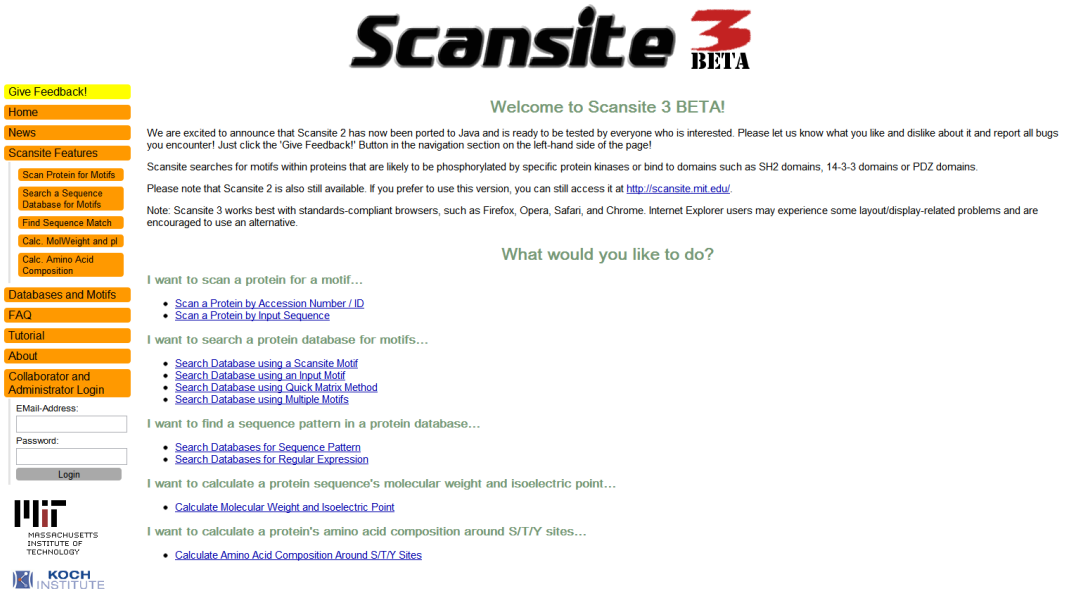


Figure 3.2: The *home*-section of Scansite 3 allows users to directly go to the search they want to run. The page is divided into three parts: The header with the Scansite-logo, the navigation section on the left-hand side, and the central content space. Some logos under the navigation bar refer to institutions that people involved in the development of Scansite (since its very beginning) were affiliated with. From here on, screenshots of Scansite 3 will focus on the content-area of the page, excluding the static header- and navigation area.

Scanning a Protein for Motifs

The key feature of Scansite is the prediction of motif-relevant sites in a given protein. This feature is referred to as *Protein Scan* or *Scan Proteins for Motifs* and allows a range of different inputs (Figure 3.3). The user has three major choices and a few minor options:

The protein to scan: Users can either choose a protein database and enter a valid identifier, or enter a protein sequence and name. Scansite 3 assists the user when entering a protein identifier by searching the selected database for identifiers that start with the letters the user entered so far (after the user entered three or more letters). This is done automatically and upon request when the *Check!*-button is clicked. If a list of protein identifiers was found, it is displayed below the textbox and if only one identifier was found, the textbox is coloured green (Figure 3.3a). The textbox turns red if no matches were found in the selected database (Fig. 3.3b). Hyperlinks to the relevant databases allow the user to quickly search for identifiers. Alternatively, users can directly paste a protein sequence (Fig. 3.3c).

The motifs to consider: It is possible to search for all motifs (Fig. 3.3a) of a motif class (mammalian or yeast motifs), for a subset of these (Fig. 3.3b), or for a user-defined motif (Fig. 3.3c). In order to use a user-defined motif, a file representing the motif's PSSM has to be prepared and uploaded. After a file of the right format was uploaded, the user is presented with an ed-

itable table that allows reviewing the motif and making some last changes before submitting. Section 3.3.1 on page 49 describes how a motif file for Scansite is created.

The stringency level: This measure defines how high sites have to score in order to be displayed as results. *High stringency* only displays the top 0.2% of sites (sites that have a score less or equal to the top 0.2% of scores in the reference proteome), *medium stringency* displays the top 1%, and *low stringency* the top 5%. Scansite 3 introduces an additional setting: *Minimum* displays the top 15%. These settings apply for motifs from the Scansite database. Since no precompiled reference-proteome analysis is available for user-defined motifs, these always display all sites with a score ≤ 5.

Additional options: The two additional options that users are given are to decide whether to show (predicted) domains in the result as supporting information, and whether to use an alternative reference proteome. If users choose to scan a protein for yeast motifs the SGD-reference proteome is selected by default (Fig. 3.3a). Domains can also be requested later on from the result-page.

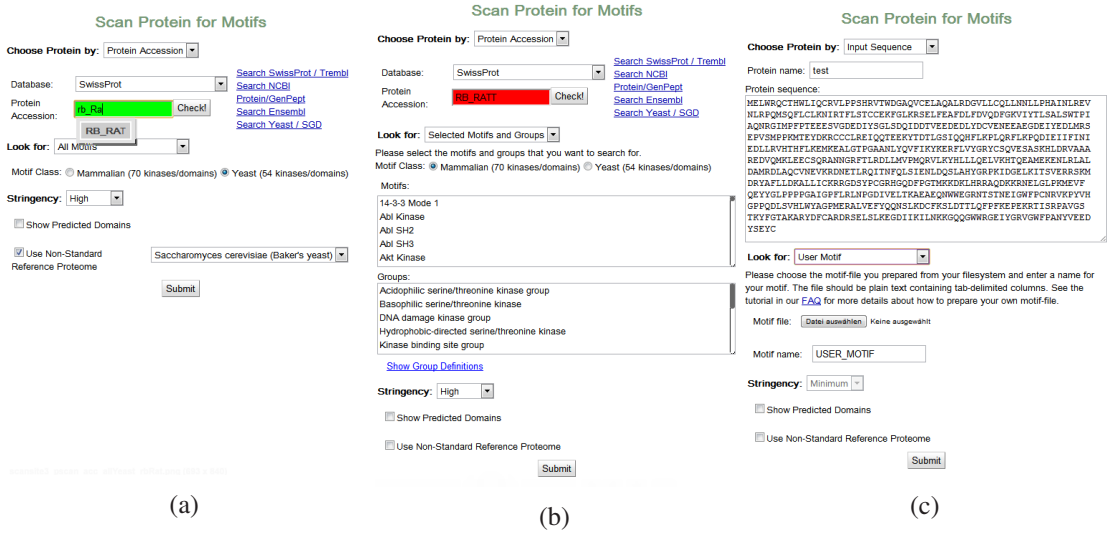


Figure 3.3: A side-by-side view of different selections in the *Protein Scan* input page showing the two options for entering a protein, and the three motif-selection options

The result-page of a high stringency protein scan for all mammalian motifs using the SwissProt protein *VAV_HUMAN* and the default reference proteome with domains is displayed in Figure 3.4. It is easy to see that the result page is split in seven sections: Protein Overview, Scan Overview, Protein Plot, Predicted Motif Sites (Table), Repeat Scan, Download Results, and Additional Analyses. Each of these sections is collapsible by clicking on the grey title-areas. This helps getting to the bottom of the page if a long list of predicted sites is displayed.

In the *Protein Overview* section, some information about the input protein is listed, including alternative identifiers and keywords (for proteins given by identifier only), and the protein's molecular

Protein Scan Results: *VAV_HUMAN (swissprot)*

Protein Overview

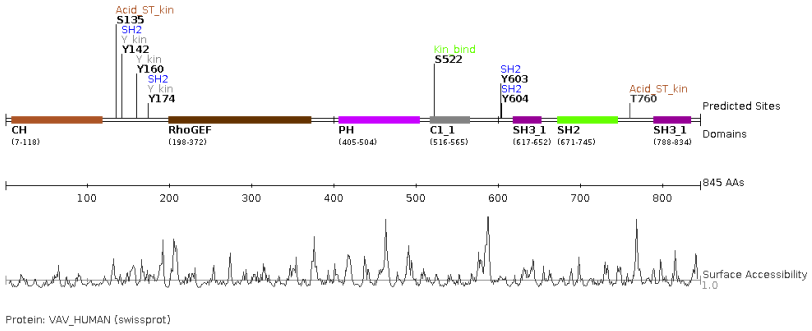
Protein Scanned: VAV_HUMAN (see SwissProt, see PhosphoSite)
Descriptions: RecName: Full=Proto-oncogene vav;
Keywords: Proto-oncogene, Polymorphism, SH3 domain, Metal-binding, Acetylation, 3D-structure, Phosphoprotein, Zinc, Zinc-finger, Reference proteome, Repeat, Complete proteome, Guanine-nucleotide releasing factor, SH2 domain
Accessions: Q15860, P15498, VAV_HUMAN
Molecular Weight: 98326.3
Isoelectric Point: 6.20

Scan Overview

Motifs searched for: 1433_m1, Abl_Kin, Abl_SH2, Abl_SH3, Akt_Kin, Amphi_SH3, AMPK, ATM_Kin, AuroA, AuroB, Cam_Kin2, Casn_Kin1, Casn_Kin2, CapC_SH3, Cdc2_Kin, CDK1_1, CDK1_2, Cdk5_Kin, Clk2_Kin, Cort_SH3, Crk_SH2, Crk_SH3, DNA_PK, EGFR_Kin, ErkDD, Erk1_Bind, Erk1_Kin, Fgr_Kin, Fgr_SH2, Fyn_SH2, Grb2_SH2, Grb2_SH3, GSK3_Kin, GSK3b, InsR_Kin, Itsn_SH3, Itk_Kin, Itk_SH2, Itk_SH3, Lck_Kin, Lck_SH2, Nck_2nd_SH3, Nck_SH2, p38_Kin, p85_SH2, p85_SH3_m1, p85_SH3_m2, PDGFR_Kin, PDK1_Bind, PDZ_nNOS_1, PDZ_nNOS_3, PDZ_class1, PDZ_class2, PIP3_PH, PKC_common, PKC_delta, PKC_epsilon, PKC_mu, PKC_zeta, PLCg_CSH2, PLCg_NSH2, PLCg_SH3, PLK1, PKA_Kin, Shc_PTB, Shc_SH2, SHIP_SH2, Src_Kin, Src_SH2, Src_SH3
Domains requested: Yes
Stringency: High
Reference histogram of all-proteome scored sites: Vertebrata (swissprot)
Number of predicted motif sites: 13

Protein Plot

Predicted PFAM-Domains (from InterProScan): CH (7 - 118), RhoGEF (198 - 372), PH (405 - 504), C1_1 (516 - 565), SH3_1 (617 - 652), SH2 (671 - 745), SH3_1 (788 - 834)
Note: The domains' positions are retrieved from InterProScan. For this reason the numbers may differ slightly from PFAM-retrieved domains.
Go to PFAM.



Predicted Motif Sites (Table)

Please allow popups in your browser settings to make links in the table work properly!

Score	Percentile	Motif	Motifgroup	Site	Sequence	Surface Accessibility	Gene Info	Previously Mapped Site
0.379	0.184%	Casein Kinase 2 (Casn_Kin2)	Acidophilic serine/threonine kinase group (Acid_ST_kin)	S135	PFPTEEEsVGDEDIY	1.1298	CSNK2B	
0.252	0.142%	PLK1 Kinase (PLK1)	Acidophilic serine/threonine kinase group (Acid_ST_kin)	T760	CFKSLDTtLQFFFKE	0.9858	PLK1	
0.166	0.017%	PDK1 Binding (PDK1_Bind)	Kinase binding site group (Kin_bind)	S522	GHDFQMFsFEETTSC	0.6835	PDPK1	
0.291	0.070%	Lck SH2 (Lck_SH2)	Src homology 2 group (SH2)	Y174	EAEGDEIyEDLMRSE	1.0397	LCK	PhosphoELM, Phosphosite
0.343	0.120%	PLCg C-terminal SH2 (PLCg_CSH2)	Src homology 2 group (SH2)	Y604	MEVFQEYyGLPPPPG	1.2297	PLCG1	
0.369	0.082%	Shc SH2 (Shc_SH2)	Src homology 2 group (SH2)	Y142	SVGDEDIySGLSDQI	0.4597	SHC1	PhosphoELM, Phosphosite
0.325	0.024%	Shc SH2 (Shc_SH2)	Src homology 2 group (SH2)	Y174	EAEGDEIyEDLMRSE	1.0397	SHC1	PhosphoELM, Phosphosite
0.399	0.179%	Shc SH2 (Shc_SH2)	Src homology 2 group (SH2)	Y603	KMEVFQEyYGLPPPP	1.3772	SHC1	Phosphosite
0.397	0.146%	Lck Kinase (Lck_Kin)	Tyrosine kinase group (Y_kin)	Y142	SVGDEDIySGLSDQI	0.4597	LCK	PhosphoELM, Phosphosite
0.412	0.199%	Lck Kinase (Lck_Kin)	Tyrosine kinase group (Y_kin)	Y160	VEEDEDLyDCVENEE	0.3286	LCK	PhosphoELM, Phosphosite
0.289	0.011%	Lck Kinase (Lck_Kin)	Tyrosine kinase group (Y_kin)	Y174	EAEGDEIyEDLMRSE	1.0397	LCK	PhosphoELM, Phosphosite
0.409	0.163%	Src Kinase (Src_Kin)	Tyrosine kinase group (Y_kin)	Y160	VEEDEDLyDCVENEE	0.3286	SRC	PhosphoELM, Phosphosite
0.312	0.022%	Src Kinase (Src_Kin)	Tyrosine kinase group (Y_kin)	Y174	EAEGDEIyEDLMRSE	1.0397	SRC	PhosphoELM, Phosphosite
Score	Percentile	Motif	Motifgroup	Site	Sequence	Surface Accessibility	Gene Info	Previously Mapped Site

DISCLAIMER: These results are purely speculative and should be used with EXTREME CAUTION because they are based on the assumption that the peptide library data is correct and sufficient to predict a site! Also, if an evidence for a site is given ('previously mapped site') it is only site- and protein-specific, meaning that this site is known to be phosphorylated by some kinase, but *not necessarily* by the kinase Scansite associates with this site!

Repeat Scan

Stringency: Medium Scan!
Repeat Search with Different Parameters

Download Results

Download results as tab separated file...

Additional Analyses

Score sites using DisPhos (Disorder-Enhanced Phosphorylation Site Predictor)...

Figure 3.4: A *Protein Scan* result page showing a number of high confidence hits, many of which are supported by PhosphoSitePlus and PhosphoELM. All the sites (except for S522) are in between functional domains predicted by InterProScan, which generally can be interpreted as a confidence-increasing fact.

weight and isoelectric point. The *Scan Overview* summarises the input-parameters of the search and displays the number of sites that have been detected using these settings. A plot of the protein gives a visual overview of the search results, annotated with some additional information in the next part of the page (*Protein Plot*). If domain information was requested, the predicted domains are listed here, followed by an image that displays the predicted sites (annotated with the position and motif group), the protein's domains (if requested) along with their names and positions, and a surface accessibility plot that shows what parts of the protein are likely to be exposed to the surface. If domains have not been requested earlier, a button will be displayed below the image that allows the user to request domain prediction at this point. The links in the list of displayed domains refer to these domains' PFAM-pages.

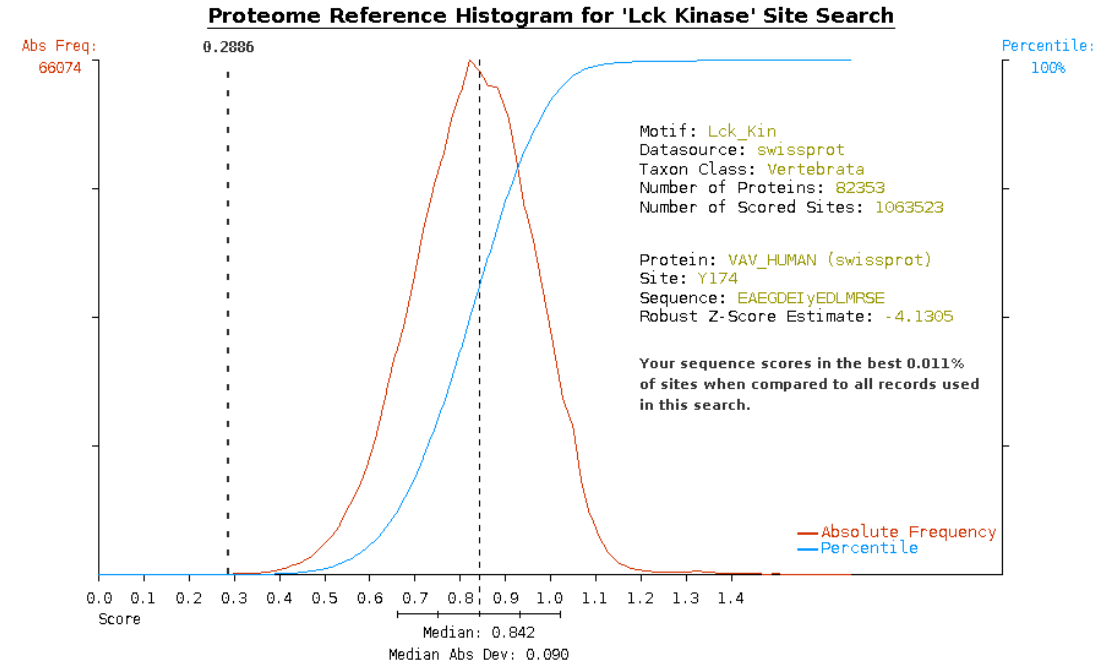


Figure 3.5: This histogram displays the whole proteome scores calculated for all vertebrate proteins in SwissProt using the motif that describes the recognition affinity of the lymphocyte-specific protein Y-kinase. In 82,353 proteins about 1 million sites were scored, the scores of which were used to draw this histogram. This histogram shows the site Tyr174 which, with a score of 0.2886, is reported to be in the top 0.011% of this histogram. The left axis shows the absolute frequency of site-scores, the right axis the percentile of scores.

The sites that are outlined in the protein plot are listed in more detail in the table view below (*Predicted Motif Sites*). All columns can be sorted by clicking on the label in the table's header. Here, each site that was found is displayed along with the motif information (motif, motif group, hyperlink to motif's gene information page), its score and percentile, and the surrounding sequence. In addition, Scansite 3 offers hyperlinks to PhosphoSite, PhosphoELM, and / or Phosida if they have reported a phosphorylation of the site before. Of course, this is only possible for proteins from

public protein databases and works best with proteins from SwissProt. The other links displayed in the table, specifically the columns *Score* and *Sequence*, refer to a histogram view of a site in the reference proteome (Fig. 3.5) and to view that shows a site's sequence highlighted in the protein's sequence (Fig. 3.6), respectively. The latter view offers a link that submits the site's sequence (15 amino acids) to NCBI's basic local alignment search tool (BLAST) (Altschul et al., 1990). This is a simple approach to see if a site is conserved in organisms that are expected to be physiologically similar to the one at hand.

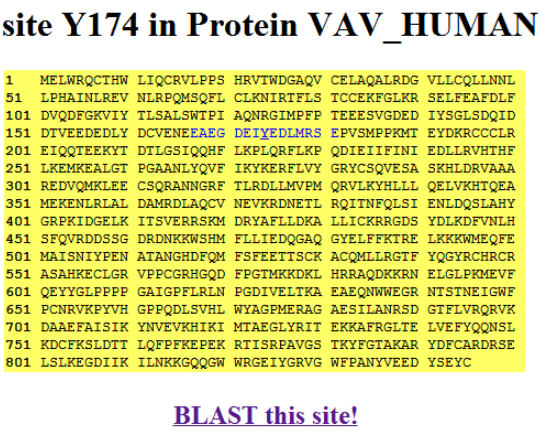


Figure 3.6: The site-in-sequence view of Tyr174 in *VAV_HUMAN*. The hyperlink at the bottom submits the site's sequence to NCBI's BLAST.

In the *Repeat Scan* section of the result page it is possible to either directly rerun the scan with a different stringency setting, or to go back to the input page to change other search parameters. The next part in the page (*Download Results*) offers a link to a downloadable version of the table shown above (tabulator-separated file). At the bottom of the result page (*Additional Analyses*) users can directly submit the current protein's sequence to DisPhos, a "Disorder-Enhanced Phosphorylation Site Predictor" (Iakoucheva et al., 2004).

Scanning a Sequence Database for Motif-Hits

The Scansite feature *Search Sequence Database for Motifs* or short *Database Search* performs the opposite search. Given a motif (or a set of motifs) and a sequence database, it searches for sequences that contain motif-relevant sites. One of the most powerful parts of this tool is the option to restrict searches to proteins of a specific organism class, species, molecular weight range, isoelectric point range, annotation, and / or sequence property. For example, this tool can be used to help identifying unknown bands in two dimensional gel electrophoresis experiments. This feature, again, allows different kinds of inputs. Here, this mainly means different ways to select motifs. Users can search for hits of single motifs, user-defined motifs, so-called "quick" motifs, and multiple motifs (Figures 3.7 and 3.8). The differences are explained in the following.

Searches for single motifs from the Scansite database are the easiest option to choose (Fig. 3.7a). If enough information about the affinities of the kinase or binding domain is known, a motif file

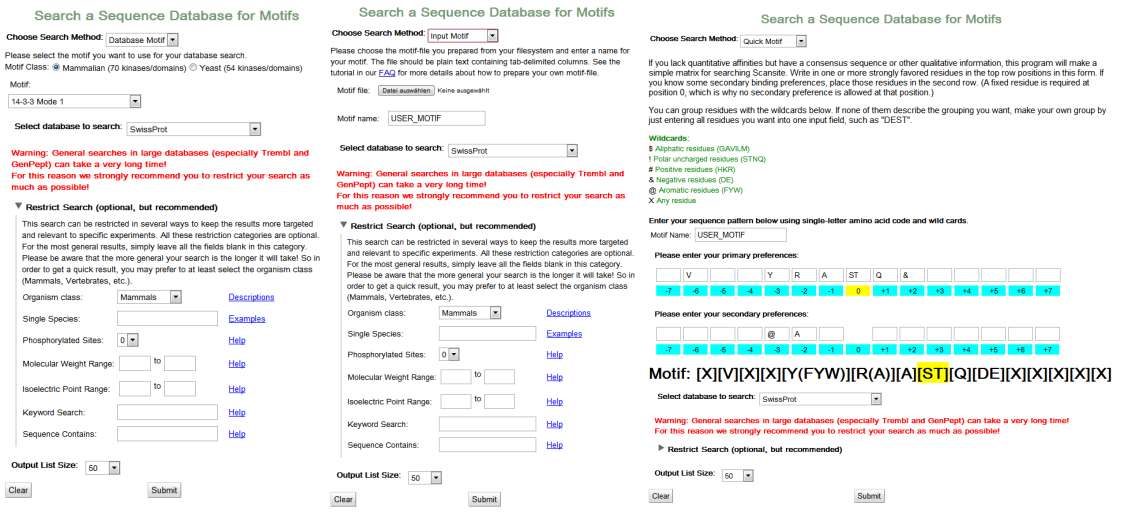


Figure 3.7: Side-by-side overview of database search input strategies that involve only one motif: a database motif, a user-defined motif, and a "quick" motif.

can be prepared, uploaded, and used for the search (Fig. 3.7b). This, however, requires a very specific idea about the motif. Often, only very little detail about a motif is known. In cases like these, the *Quick Motif* option is the best choice (Fig. 3.7c). For defining a quick motif, the user can enter a set of primary and secondary preferences which are then turned into a Scansite motif. The web-page describes a number of wildcards that can be used to easily define preferences for, e. g., hydrophobic or positive residues. It is also possible to search for sequences that not only contain one, but up to five motifs (Fig. 3.8). These searches can include either database-motifs, user-defined motifs, or a combination of both. The score for multi-motif sites is the mean of the scores of the sites involved. Co-occurrences of different motif-sites in proteins can be restricted and filtered in different ways: First of all, it is possible to penalise gaps between sites of different motifs (Fig. 3.8a). Gap penalty settings are either *high*, *medium*, *low*, or *none*. Penalties p are then added to the score according to the maximum distance d_{max} between the involved sites (i. e. position of site closest to C-terminus minus position of site closest to N-terminus), resulting in the following penalty-values: $p_{low} = 0.001 \cdot d_{max}$; $p_{medium} = 0.01 \cdot d_{max}$; $p_{high} = 0.1 \cdot d_{max}$. Secondly, it is possible to define strict minimum and maximum distance bounds between motif-specific sites (Fig. 3.8b). For example, a search for the motifs of the DNA damage kinases *DNA PK* and *ATM*, and *P38 MAPK* can be restricted in the following way: sites of *DNA PK* and *ATM* have to be at least 10 residues and at most 90 residues apart, and sites of *ATM* and *P38 MAPK* are required to be closer than 150 residues. The number of restrictions that can be defined is limited to three. This example is also shown in Figure 3.8b.

Since *Database Searches* may find a very high number of results, the number of sites that are displayed is limited. By default, the size of the output list is limited to 50, but users can also choose the sizes 100, 200, 500, 1000, and 2000. This is just the number of sites that are displayed in the table on the result page. A file containing all the hits that were found in this search can be

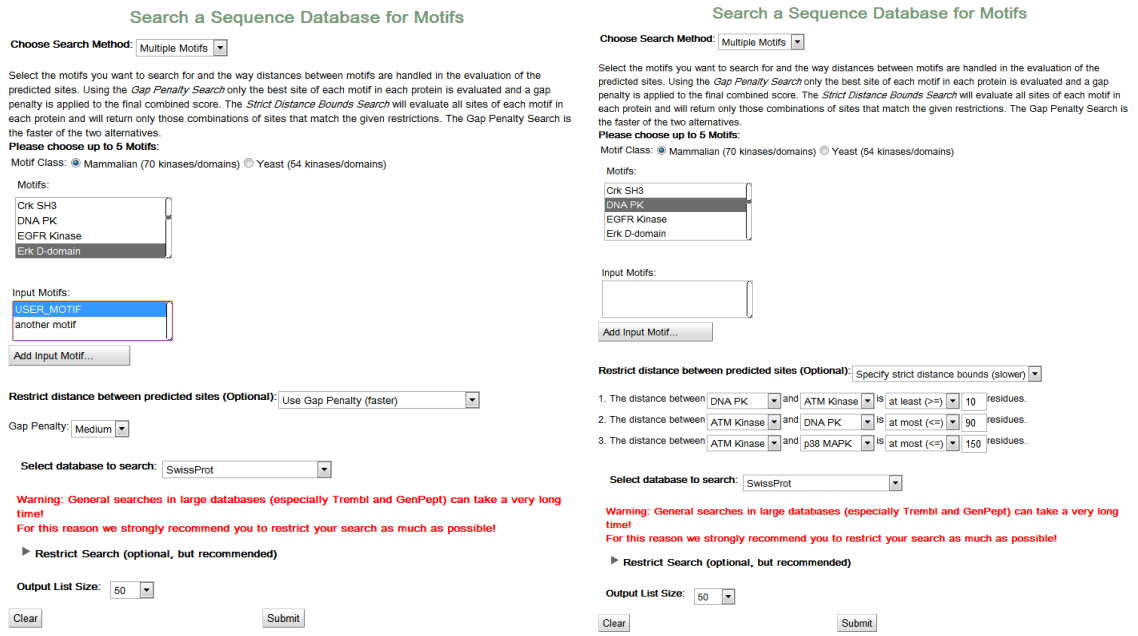


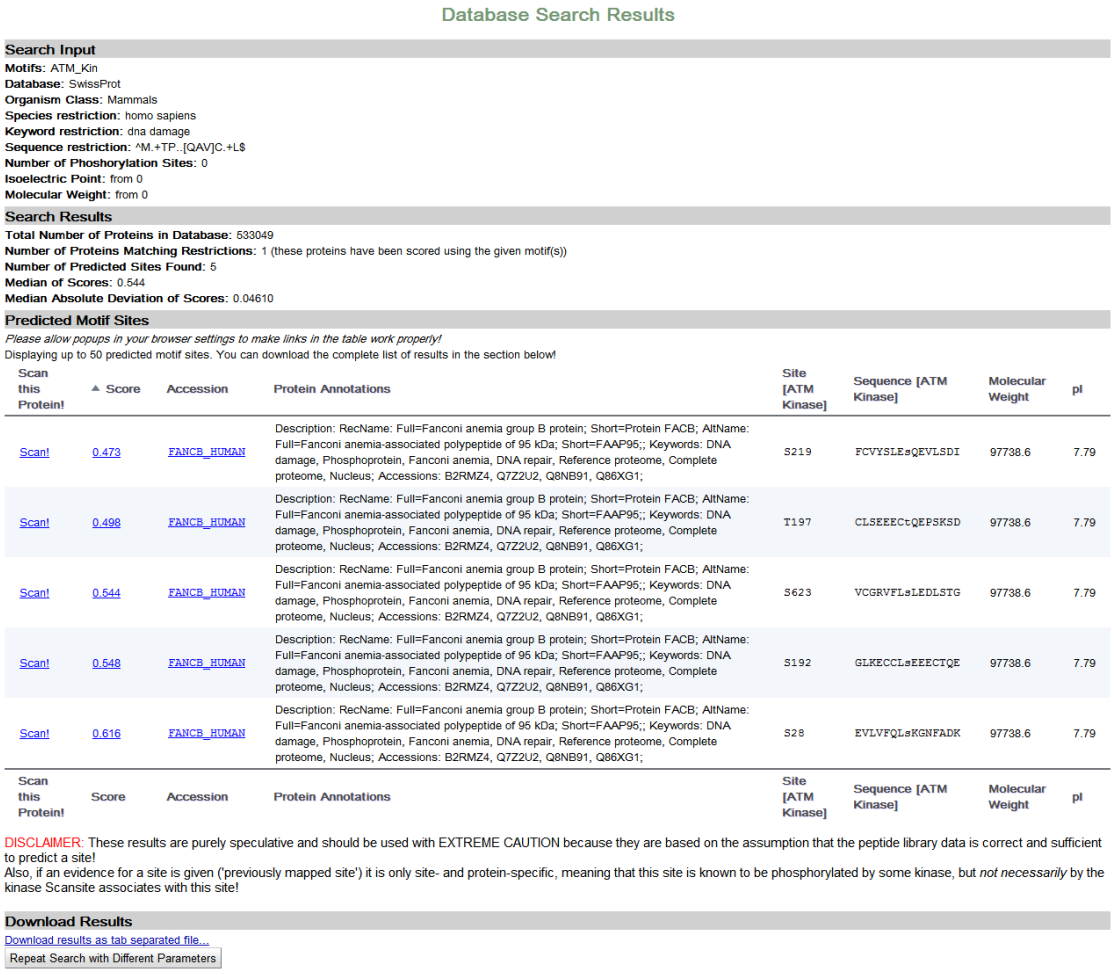
Figure 3.8: Both multiple motif search options. The gap penalty setting on the left, the strict limit option on the right. Two user-defined motifs were uploaded in the gap penalty version.

downloaded too.

A result page of a *Database Search* is displayed in Figure 3.9. Here, four sections can be distinguished: The *Search Input* part of the page summarises the preferences defined in the input page. *Search Results* gives an overview of the number of proteins in the entire sequence database, the number of proteins found that match the given restrictions, and the number of sites found in these proteins. In addition, the median and MAD of these sites' scores is displayed. This part is followed by a table-view of the sites found (*Predicted Motif Sites*). The table shows the (combined) site score, some information about the protein that was found (including MW and pI), and displays some site-specific information (site and surrounding sequence). For multi-motif searches a site and sequence column for each motif in the motif's site is given. The first column in the table allows to directly scan the protein for other motifs. The link in the column labelled *Accession* takes the user to the protein's page in its primary database. The score-column links to a histogram that shows the site's score in comparison to all scores found in that search. At the bottom of the page options for downloading the entire result-set and for repeating the search are given.

Searching Sequence Database for Simple Patterns

Scansite can find strictly defined sequence patterns in a protein databases by searching for a sequence pattern defined by a regular expression. Although this is the most simple search feature, it is also very powerful, especially because it allows users to search for sequences that contain one



Database Search Results

Search Input

Motifs: ATM_Kin
Database: SwissProt
Organism Class: Mammals
Species restriction: homo sapiens
Keyword restriction: dna damage
Sequence restriction: ^M.+TP..[QAV]C.+L$
Number of Phoshorylation Sites: 0
Isoelectric Point: from 0
Molecular Weight: from 0

Search Results

Total Number of Proteins in Database: 533049
Number of Proteins Matching Restrictions: 1 (these proteins have been scored using the given motif(s))
Number of Predicted Sites Found: 5
Median of Scores: 0.544
Median Absolute Deviation of Scores: 0.04610

Predicted Motif Sites

Please allow popups in your browser settings to make links in the table work properly!
Displaying up to 50 predicted motif sites. You can download the complete list of results in the section below!

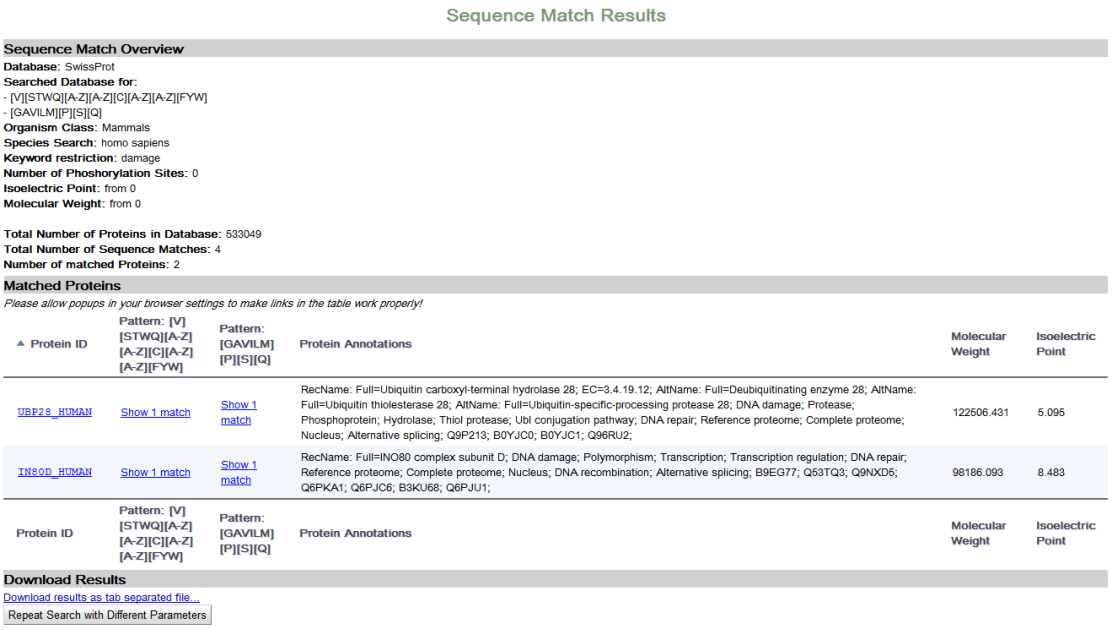
Scan this Protein!	▲ Score	Accession	Protein Annotations	Site [ATM Kinase]	Sequence [ATM Kinase]	Molecular Weight	pI
Scan!	0.473	FANCB_HUMAN	Description: RecName: Full=Fanconi anemia group B protein; Short=Protein FACB; AltName: Full=Fanconi anemia-associated polypeptide of 95 kDa; Short=FAAP95;; Keywords: DNA damage, Phosphoprotein, Fanconi anemia, DNA repair, Reference proteome, Complete proteome, Nucleus; Accessions: B2RMZ4, Q7Z2U2, Q8NB91, Q86XG1;	S219	FCVYSLEsQEVLSDI	97738.6	7.79
Scan!	0.498	FANCB_HUMAN	Description: RecName: Full=Fanconi anemia group B protein; Short=Protein FACB; AltName: Full=Fanconi anemia-associated polypeptide of 95 kDa; Short=FAAP95;; Keywords: DNA damage, Phosphoprotein, Fanconi anemia, DNA repair, Reference proteome, Complete proteome, Nucleus; Accessions: B2RMZ4, Q7Z2U2, Q8NB91, Q86XG1;	T197	CLSEEECtQEPSKSD	97738.6	7.79
Scan!	0.544	FANCB_HUMAN	Description: RecName: Full=Fanconi anemia group B protein; Short=Protein FACB; AltName: Full=Fanconi anemia-associated polypeptide of 95 kDa; Short=FAAP95;; Keywords: DNA damage, Phosphoprotein, Fanconi anemia, DNA repair, Reference proteome, Complete proteome, Nucleus; Accessions: B2RMZ4, Q7Z2U2, Q8NB91, Q86XG1;	S623	VCGRVFLsLEDLSTG	97738.6	7.79
Scan!	0.548	FANCB_HUMAN	Description: RecName: Full=Fanconi anemia group B protein; Short=Protein FACB; AltName: Full=Fanconi anemia-associated polypeptide of 95 kDa; Short=FAAP95;; Keywords: DNA damage, Phosphoprotein, Fanconi anemia, DNA repair, Reference proteome, Complete proteome, Nucleus; Accessions: B2RMZ4, Q7Z2U2, Q8NB91, Q86XG1;	S192	GLKECCLsEEECTQE	97738.6	7.79
Scan!	0.616	FANCB_HUMAN	Description: RecName: Full=Fanconi anemia group B protein; Short=Protein FACB; AltName: Full=Fanconi anemia-associated polypeptide of 95 kDa; Short=FAAP95;; Keywords: DNA damage, Phosphoprotein, Fanconi anemia, DNA repair, Reference proteome, Complete proteome, Nucleus; Accessions: B2RMZ4, Q7Z2U2, Q8NB91, Q86XG1;	S28	EVLVFQLsKGNFADK	97738.6	7.79
Scan this Protein!	Score	Accession	Protein Annotations	Site [ATM Kinase]	Sequence [ATM Kinase]	Molecular Weight	pI

DISCLAIMER: These results are purely speculative and should be used with EXTREME CAUTION because they are based on the assumption that the peptide library data is correct and sufficient to predict a site!
Also, if an evidence for a site is given ('previously mapped site') it is only site- and protein-specific, meaning that this site is known to be phosphorylated by some kinase, but *not necessarily* by the kinase Scansite associates with this site!

Download Results

Download results as tab separated file...
Repeat Search with Different Parameters

Figure 3.9: The result of a very rigorously restricted database search using *ATM*: Only one protein (*FANCB_HUMAN*) was found that matches the given restrictions. Within this protein 5 putative *ATM* phosphorylation sites were predicted.



Sequence Match Results

Sequence Match Overview

Database: SwissProt
Searched Database for:
- [V][STWQ][A-Z][A-Z][C][A-Z][A-Z][FYW]
- [GAVILM][P][S][Q]
Organism Class: Mammals
Species Search: homo sapiens
Keyword restriction: damage
Number of Phoshorylation Sites: 0
Isoelectric Point: from 0
Molecular Weight: from 0

Total Number of Proteins in Database: 533049
Total Number of Sequence Matches: 4
Number of matched Proteins: 2

Matched Proteins

Please allow popups in your browser settings to make links in the table work properly!

▲ Protein ID	Pattern: [V][STWQ][A-Z][A-Z][C][A-Z][A-Z][FYW]	Pattern: [GAVILM][P][S][Q]	Protein Annotations	Molecular Weight	Isoelectric Point
UBP28_HUMAN	Show 1 match	Show 1 match	RecName: Full=Ubiquitin carboxyl-terminal hydrolase 28; EC=3.4.19.12; AltName: Full=Deubiquitinating enzyme 28; AltName: Full=Ubiquitin thiolesterase 28; AltName: Full=Ubiquitin-specific-processing protease 28; DNA damage; Protease; Phosphoprotein; Hydrolase; Thiol protease; Ubl conjugation pathway; DNA repair; Reference proteome; Complete proteome; Nucleus; Alternative splicing; Q9P213; B0YJC0; B0YJC1; Q96RU2;	122506.431	5.095
IN80D_HUMAN	Show 1 match	Show 1 match	RecName: Full=INO80 complex subunit D; DNA damage; Polymorphism; Transcription; Transcription regulation; DNA repair; Reference proteome; Complete proteome; Nucleus; DNA recombination; Alternative splicing; B9EG77; Q53TQ3; Q9NXD5; Q6PKA1; Q6PJC6; B3KU68; Q6PJU1;	98186.093	8.483
Protein ID	Pattern: [V][STWQ][A-Z][A-Z][C][A-Z][A-Z][FYW]	Pattern: [GAVILM][P][S][Q]	Protein Annotations	Molecular Weight	Isoelectric Point

Download Results

Download results as tab separated file...

Repeat Search with Different Parameters

Figure 3.10: The *Sequence Match* result page of a search for two patterns which were found once in each of the two proteins returned by the search, totalling four pattern matches

or more (up to five) simple amino acid patterns. In addition, it is easy to create statistics about occurrences of sequence patterns (e. g. in different organisms) in protein databases. This helps to find out how common or uncommon a pattern is. The feature that allows users to do searches like these is referred to as *Find Sequence Match*, or simply *Sequence Match*. Users can directly search for a regular expression, or have Scansite create regular expressions from sequence-preferences. Entering a regular expression allows a very flexible search that at the same time can be very specific. Entering sequence patterns offers the option of searching for multiple patterns at once. Also, the latter alternative assists the user in creating a search pattern, so that this feature can easily be used even if the user does not know anything about regular expressions. Figure 3.10 shows a sequence match result page, displaying the two proteins that each contain the two queried patterns once. The table displays information about the matched proteins and provides links to the protein's page in its native database. In addition, links to the matched sites in the sequence are provided. At the bottom of the page, a tabulator-separated file that contains this search's entire result-set can be downloaded.

Other features

In addition to the search options described before, Scansite also provides a feature that calculates and displays the molecular weight and isoelectric point for protein sequences with a user-defined number of hypothetical phosphorylation sites (*Calculate MolWeight and pI* in navigation section).

Another helpful feature can be used to *Calculate the Amino Acid Composition* in a given protein around a given amino acid. Once a protein (either by sequence or by accession) is submitted, a matrix is displayed that shows the composition of amino acids around the selected centre residue in

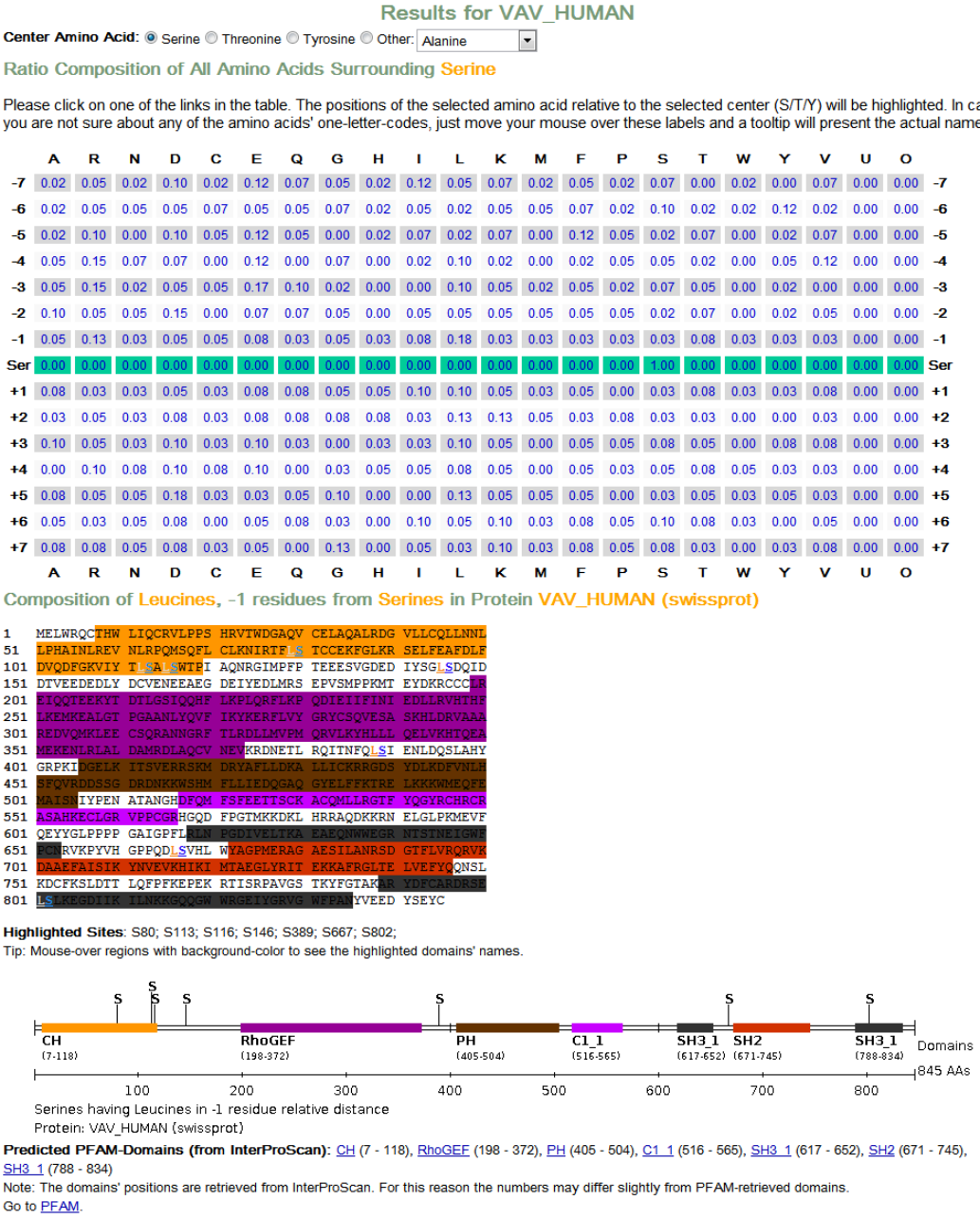


Figure 3.11: The amino acid composition view of the protein *VAV_HUMAN*. The table shows the relative frequencies of amino acids surrounding the centre residue, serine. Lysine at position -1 has been selected in this view, and the sequence and image view below show all occurrences of serines preceded by lysines in the given sequence. In addition, the protein's domains are displayed in both these views.

relative numbers (Figure 3.11). By default, the centre residue is serine, but all other amino acids can be selected as well. The matrix that is shown displays the relative frequencies of other amino acids ± 7 residues from the selected centre. Hence, the matrix has 22 columns (amino acids) and 15 rows (positions around the centre residue). For example, the value of 0.13 in position -1 for arginine (R) displayed in the matrix in Figure 3.11 means that out of all amino acids in this sequence that directly precede serines (centre) 13% are arginines. In the screenshot, leucine in position -1 is selected, showing 18%. The values in the matrix are rounded, so not all row-values sum up to exactly 1. When a protein is submitted for the composition overview, InterProScan is used to search for domains in the sequence. A message is displayed when the domains are ready to be displayed.

All values in the matrix are clickable. Clicking on a value highlights the selected residue in the two views below the matrix: the sequence view and the image-view (which is shown as soon as the domain search has finished). Wherever the selected residues (centre and selected) occur in the chosen distance from each other, the sequence view highlights both residues. The image only shows the centre residue in order to not clutter the image too much. If domains were found, they are displayed in both views in matching colours.

Restricting Database and Sequence Searches

All database searches (including searches using the sequence match feature) can and, most importantly, *are recommended to be* restricted in some way. For most use-cases it is not necessary to search an entire protein database, since the (hopefully targeted) question driving the search usually provides enough prior knowledge to exclude a large group of proteins by, for example, focusing on a single species. Also, the broader a search is (i. e. the more proteins are included in the search), the longer the search takes, and the more irrelevant sites are included in the list of results. This again, makes examining the resulting predictions harder. The main reason why this is of relevance is the enormous size of many protein databases that are included in Scansite 3, especially GenPept and TrEMBL which include around 9 and 18 million protein entries, respectively. Complete searches in these two resources can take up to hours. This is why it is recommended to restrict searches. All the restrictions utilise the power of filtering strategies in queries to relational databases. This handling not only makes the filtering process much easier from a programming point-of-view, but also saves time since a smaller result-set is transferred to from the database to the application. However, in order to make it possible to perform queries like that based on the Scansite-relevant data, it is necessary to save data in a redundant way and to store values that could be calculated from others (e. g. isoelectric points and molecular weights can be calculated from the sequence, but have to be stored in order to allow efficient queries).

Restrictions can be applied in different ways: To start with, the search can be restricted to proteins from a specific species or organism class. Species can be entered using regular expressions to match not only a single species, but a number of related species. For example, all species of genus *Homo* can be found by entering “^Homo”. An organism class in Scansite does not directly correspond to the taxonomic definition of a *class*, but is a term for groups of species that are frequently used, such as “Bacteria”. The following organism classes are defined in Scansite:

Mammals: Members of the actual taxonomic class *Mammalia*.

Vertebrates: Members of the sub-subphylum *Vertebrata*, including class *Mammalia*.

Invertebrates: All members of superkingdom *Eukaryota* excluding vertebrates, plants, and fungi.

Plants: Members of kingdom *Viridiplantae*.

Fungi: Members of kingdom *Fungi*.

Bacteria: Members of superkingdoms *Archaea* and *Eubacteria*.

Viruses: Members of superkingdom *Viridae*.

Other: Members that don't fit in any other category, for example, plasmids and synthetic sequences.

All: All organisms, including those from category *Other*.

Next, upper and lower boundaries can be defined to restrict the search to proteins of certain weight (MW, in Dalton) or isoelectric point (pI), which is calculated according to Bjellqvist et al. (1993). As these values vary dependent on whether a protein is phosphorylated (and if so, how many times) or not, these boundaries can be applied to a number between zero and three hypothetical phosphorylations. In addition, if the search is performed in a well-annotated protein database like SwissProt, a keyword can be included to restrict the search to proteins that are annotated with this keyword. This can be used to find proteins annotated as, for example, phosphatases. The database search (not, however, the sequence match search) also includes the option of adding a sequence match requirement. This means, that only proteins are included in the search that match a given sequence pattern. This can be used to apply a motif search to a previously run sequence match search.

Creating a Scansite Motif

Both main search options in Scansite allow the use of user-defined motifs. These motifs have to be defined in tabulator-separated text files (commonly known as tsv-files) of a Scansite-specific format. Here, the most important steps to create a file like this are described. All user-defined motifs that are uploaded to Scansite 3 are only used for the user's searches and are deleted as soon as the user leaves the site.

The main idea is to create a file that represents a motif's PSSM with values that Scansite's scoring algorithm can deal with. PSSMs in Scansite describe amino acid specific affinity values for a sequence window of 15 residues. Lines correspond to positions in the sequence window, columns (separated by tabulators) to amino acids. The first line (row 1) defines the residue-to-column assignments by the amino acids' one letter codes. In order to let Scansite know which residue is the deciding one for a given PSSM, Scansite requires (at least) one residue to be invariant in the motif sequence. For example, the fixed residue should be a tyrosine for motifs recognized by

tyrosine-kinases, and serine and threonine for S-/T-kinases. The middle position, row 9, holds the fixed residue in the matrix which is defined by a value of 21. This value is intended to be used only for fixed residues. The N-terminal side of the motif is defined by rows 2 through 8, and rows 10 through 16 hold the scores for the C-terminal side.

The number of columns in the matrix can vary: If, for example, the underlying peptide library screen did not include some residues, these columns can be left out of the matrix and Scansite will assign them default values of 1 everywhere except in the fixed position, where the default score is 0. Scansite can also deal with special requirements: A motif's preference for a protein sequence's N- or C-terminus can be incorporated by using a column labelled "$" (dollar sign) or "*" (asterisk), respectively. This applies, for example, to PDZ-domain recognising motifs which recognise C-terminal regions. These positions are assigned values of 0 by default. As mentioned before, Scansite 3 also recognises the rarely occurring amino acids selenocysteine (U) and pyrrolysine (O), which can be added by their one letter code as well. Due to their similar chemical structure, the default numbers for these residues are the values of cysteines and lysines, respectively. Lastly, some wildcard values can be used for very special cases: "B" (aspartate/asparagine), "Z" (glutamate/glutamine), "J" (leucine/isoleucine), and "X" (any residue). These symbols are included because they occur rarely in public protein databases, but they generally have no relevance for actual research purposes. The default values for these wildcards are the mean values of the amino acids that they encode. Therefore, a user-defined Scansite motif has 16 rows (15 rows of numbers + one header) and between 1 (the fixed residue) and 28 (including wild card amino acids) residues. A sample matrix is shown in Table 3.1.

Scansite's scoring system ranges from 0 to roughly 21. Giving an individual amino acid a score of 1 at one position in the motif indicates that no preference exists, positive or negative, for that particular amino acid in that position. Giving all amino acids in one position of the motif a score of 1 (i. e. making all values in a single row of the matrix equal to one) indicates no preference exists for any particular residue type at that position in the motif. Values higher than 21 are permitted to indicate very strong affinities. However, negative values are not permitted for defining a strong disfavouring of amino acids. Instead, values between zero and one should be used for that purpose. Beware that the scoring function uses natural logarithms, so values less than 1, particularly those less than 0.5, strongly penalise for that particular residue in a motif. In fact, the penalty of negative selection from a matrix value of 0.1 in the final score is equivalent (though opposite) to the positive selection obtained with a value of 10 for another residue in the motif.

3.3.2 Using the Web Service

The Scansite 3 web service[11] is intended to be used for batch-processing motifs and proteins. However, not all features are available as a web service, but only those that are expected to be used in that manner. The features that are offered for computational access are single-motif *Database Searches* with Scansite-motifs, *Sequence Match* searches using regular expressions, and *Protein*

[11] *http://scansite3.mit.edu/Scansite3Webservice*

Table 3.1: An example of what a user-defined Scansite motif file may look like. To start with, the order of the columns does not matter. This example shows a PSSM that encodes a motif with central tyrosine (Y with central value 21). Each row shows affinity values for amino acids in positions around this central residue.

A Sample Scansite Motif

N	P	Y	F	G	A	C	Q	R
1.12	0.6	1	0.4	1	1.2	1	2.125	1
5.278	1	2.2	0.413	1	2	1.6	0.98	1
0.11	0.872	0	2.89	0.6	2.43	0.88	0.298	1
4.28	0.62	0	1.124	3.19	1.78	1.372	0.411	1
3.13	0.451	0	1.35	3.882	0.724	0.736	0.62	1
1.789	0.62	0	0.88	3.55	0.54	1.07	0.451	2.12
3.123	0.548	0	1.29	1.45	1.151	1.155	0.62	1
0	0	21	0	0	0	0	0	0
2.0	0.411	0	2.93	0.76	1.07	0.357	0.11	1.21
1.91	0.12	0	1.36	0.97	1.26	2.219	4.28	2.2
8.21	0.14	0	2.52	0.38	0.686	0.994	3.13	0.2
1.02	0.5	2.2	0.61	0.68	0.781	0.895	3.19	1.2
0.213	2.12	0	1.29	2.13	1	1	6.88	1
0.82	1.6	0.2	0.4	1.10	0.98	2	1	1
0.11	1	0	0.4	1	0.298	3.12	1.21	1

Scans. In addition, some utility functions allow access to Scansite data, including queries for all available motifs and datasources. All of these can be accessed in a RESTful service-like manner. Generally, the features can be accessed with URIs of this format:

```
http://[WEBSERVICE-URI]/[FEATURE](/[PARAMETER=(VALUE)?])*
```

This annotation means that after the web service's URI (uniform resource identifier) an identifier of the feature that the user wants to use follows, separated by a slash ("/"). This can be followed by a number of key-value pairs that represent parameters and their values, each of which is separated from the rest of the URI by a slash. The question mark ("?") and the asterisk ("*") represent optional values that may occur once or multiple times, respectively. Parentheses are used to group the scope of these quantifiers: For some features one or more parameters are required, some features require parameters, but have optional values, and others do not require any parameters at all. The result of all web service features are valid and well-formed XML-files with self-explanatory tag-names. These can easily be parsed and processed using state-of-the-art XML-parsers that are available for all commonly used programming and scripting languages.

Some of the parameter-options in the features described below allow only a restricted set of values, which can be acquired by using the utility-functions that are provided. Table 3.2 shows a list of abbreviations used in the definition of all the services. In the following the web service's base URI

Table 3.2: Overview of the abbreviations and quantifiers used in the description of the web service's features. The services mentioned in the descriptions listed here are utility services that return all possible values for a given parameter-option or give information about whether an entity exists in the Scansite 3 database. For example, the *motifDefinitions*-service returns information about all the motifs that are currently publicly available and the *proteinExists*-service returns *true* or *false* dependent on whether a given protein identifier was found in the given database.

Abbreviations used in web service definitions	
[URI]	The web service's URI.
[ANY]	Any value
[DEC]	Numbers with decimal point are allowed
[NUM]	Integer value
[NP]	Only a number in the range [0–3] is allowed
[MC]	Only a motif class as returned by the *motifClasses*-service is permitted
[M]	Only one motif nickname as returned by the *motifDefinitions*-service is permitted
[MS]	Only motif nicknames as returned by the *motifDefinitions*-service are permitted. If multiple motifs are used, they have to be separated by a tilde ("∼")
[DS]	Only datasource's nicknames as returned by the *datasources*-service are permitted
[OC]	Only organism classes as returned by the *organismClasses*-service are permitted.
[ST]	Only stringency values as returned by the *stringencyValues*-service are permitted
[P]	A valid protein identifier is required. The *proteinExists*-service can be used to find out whether a protein exists in Scansite's database. This service returns true for *valid* protein identifiers.
[SEQ]	Only a protein sequence (i. e. amino acid one letter codes) is permitted
?	Optional parameter: The value of the parameter (i. e. the right-hand-side of the equals-sign "=") can be left blank. In general, parameters are mandatory. Only parameters with this quantifier are optional!

is excluded from the service-string in order to save space. For a similar reason, line breaks are introduced that are not part of the service-URIs. These are indicated by a backslash ("\").

Datasources: Get a list of all datasources that are available. The datasources' nicknames in the returning XML-file can be used for defining a datasource in other services.

```
[URI]/datasources
```

Organism Classes: Gets a list of all the organism classes that can be used to restrict a database search or sequence match.

```
[URI]/organismClasses
```

Motif Classes: This service returns a list of all motif classes that are available at the moment.

```
[URI]/motifClasses
```

Motifs: Service for getting information about all the motifs associated with the given class.

```
[URI]/motifDefinitions/motifClass=[MC]
```

Stringency Values: Returns a list of valid stringency options.

```
[URI]/stringencyValues
```

Proteins: Service for checking if a protein identifier is found in one of Scansite's data-repositories. It requires a protein identifier and a datasource nickname as parameters.

```
[URI]/proteinExists/accession=[ANY]/datasourceNickname=[DS]
```

Protein Scan with Protein Identifier: Runs a *Protein Scan* with the given parameters on the given protein. Before this service is started, it is recommended to check the protein's availability in the database with the *proteinExists*-service. The parameter *motifNicknames* is marked as optional. This means that, if no motif is given, all motifs of the given motif class are included in the scan.

```
[URI]/proteinScan/accession=[P]/datasourceNickname=[DS]/motifClass=[MC] \
     /motifNicknames=[MS]?/stringencyValue=[ST]
```

Protein Scan with Protein Sequences: Runs a *Protein Scan* with an input sequence and the given scan-parameters. As for the *proteinScan*-feature described above, the absence of motif-definitions means that all motifs of the given class are included in the search.

```
[URI]/proteinScan/proteinName=[ANY]?/sequence=[SEQ]/motifClass=[MC] \
     /motifNicknames=[MS]?/stringencyValue=[ST]
```

Sequence Match: Runs a *Sequence Match* search using a regular expression pattern and the defined restrictions.

```
[URI]/sequenceMatch/sequenceMatchRegex=[ANY]/datasourceNickname=[DS] \
     /organismClass=[OC]/speciesRestrictionRegex=[ANY]? \
     /numberOfPhosphorylations=[NP] \
     /molWeightFrom=[NUM]?/molWeightTo=[NUM]? \
     /isoelectricPointFrom=[DEC]?/isoelectricPointTo=[DEC]? \
     /keywordRestrictionRegex=[ANY]?
```

Database Search: A service for running a *Database Search* using a Scansite motif of a given class and a set of restrictions.

```
[URI]/databaseSearch/motifNickname=[M]/datasourceNickname=[DS] \
     /organismClass=[OC]/speciesRestrictionRegex=[ANY]? \
     /numberOfPhosphorylations=[NP] \
     /molWeightFrom=[NUM]?/molWeightTo=[NUM]? \
     /isoelectricPointFrom=[DEC]?/isoelectricPointTo=[DEC]? \
     /keywordRestrictionRegex=[ANY]?/sequenceRestrictionRegex=[ANY]?
```

The following example will demonstrate how these services are intended to be used: In order to repeat the *Protein Scan* described in Section 3.3.1 using a web service the user has to first check if the protein exists in the database of interest, and what the datasource's identifier ("nickname") is. In order to do this, the *datasources* service is used, which returns the nickname "swissprot" for

the SwissProt database. This can then be used to check if the protein of interest *VAV_HUMAN* is available in Scansite's mirror of SwissProt.

```
[URI]/proteinExists/accession=vav_human/datasourceNickname=swissprot
```

If this service returns *false*, either the given protein accession is invalid, or it does not exist in Scansite's database. Here, the service returns *true*, which means it is safe to continue. Next, the rest of the parameters for the protein scan have to be defined: This search should use the highest stringency value. In order to determine the key to use, the *stringencyValues*-service is used. It returns "High" as the key to use. The *motifClasses*-service offers "YEAST" and "MAMMALIAN" motifs. Here, the latter is used. Since the search should include all motifs of this class, no information about motif's identifiers is needed. This results in the following query URI:

```
[URI]/proteinScan/accession=VAV_HUMAN/datasourceNickname=SWISSPROT \
     /motifClass=MAMMALIAN/motifNicknames=/stringencyValue=High
```

The result file contains a collection of sites, information about these sites, and the protein's sequence and identifier. All other features included in the web service can be used in a similar manner.

3.4 Technical aspects of Scansite 3

The main goal in implementing Scansite 3 was to re-engineer Scansite 2 from scratch using state-of-the art web-technologies to create a software framework that can easily be extended, debugged, and updated. Another important goal in this project was to create an easy-to-use web-interface that not only is easy to use for Scansite 2 users, but also intuitive for people who have not worked with Scansite before.

Scansite 3 was completely implemented in Java using the Google Web Toolkit library[12] for the client-side web interface. Additional libraries were used for client-server communication (GWT Dispatch[13]) and file-upload (GwtUpload[14]). Using a single programming language for both the server-side and the client-side code makes the development of new features much easier. Of course, the code that defines the user-interface cannot be interpreted and displayed by browsers. Instead, the GWT compiler creates highly optimised JavaScript code from the Java-code which browsers can deal with easily. This is a remarkable advantage, since it is not straight-forward to (1) integrate Java and JavaScript code in one application and (2) to do this in a performant way. Hence, a lot of issues can be avoided this way. The web service uses the Jersey library[15], Oracle's reference implementation of JAX-RS (Java API for RESTful Web Services). The data-backend is a MySQL database that, generally speaking, stores motif- and substrate-related information and

[12] *http://developers.google.com/web-toolkit/*
[13] *http://code.google.com/p/gwt-dispatch/*
[14] *http://code.google.com/p/gwtupload/*
[15] *http://jersey.java.net/*

mirrors several publicly accessible protein databases for performance reasons. This design allows easy additions of new public databases or other protein collections, like domain-specific subsets of databases, and the easy addition of new motifs. Some tables needed to be denormalised in some way in order to allow more performant access. Although Scansite 3 was only tested with Apache Tomcat so far, it can be deployed in any web application server. In addition to the user-accessible web-front end, a number of Java command line applications have been created that can be used to populate and update the data-backend.

Chapter 4

Enrichment of well-known Protein-Protein Interaction Networks with Novel and Relevant Interactors

Single healthy proteins (i. e. correctly transcribed from non-mutated genes and properly translated) do exactly what they are biochemically capable of, which is what they are — from an evolutionary point of view — supposed to do. They do not know what their part in a bigger biological picture (e. g. that of its organism) is. Kinases assist in the phosphorylation of proteins, but do not care what their substrates are going to do once they are phosphorylated. Each of these entities is restricted to this very constrained view of its surroundings. This very view is the same scope that all computational protein-protein interactionprediction tools, including Scansite 3, have. From a semantic point of view, much more insight and information can be found by focusing not only on single interactions, but on the biological context that the interactions happen in. For example, a single phosphorylation event might induce other phosphorylation events and so on, which in further consequence may result in a given phenotype. This view outlines the *cellular signalling pathway view* of biological systems that has been used to describe many important biological phenomena (e. g. blood coagulation pathways or insulin signal transduction pathway). Connecting pathways at points where they share genes allows an even more relevant and closer-to-complete view of the actual interaction network. Merging pathway information with independently reported binary interaction data resulted in a variety of public protein-protein interaction databases (some of which are described in Section 2.1.2 on page 16). However, the interaction networks that are currently available in these databases are probably far from complete and may include a large number of false positives as they store data originating from different types of experiments executed by different research groups. There are many caveats associated with the data stored in these databases, with the origin of the data being the most important one. Also, these networks are too complex and big to be analysed manually. Luckily, some databases provide scores that give some estimate about how likely interactions are to be true. Although these scores are just a simple predictor of the confidence of interactions, they are very helpful as they allow some comparison of sometimes confusingly

annotated interaction data, which makes it possible to distinguish interactions that are more likely to be true from others.

Parts of the reported interactome represent the accepted standard of what the scientific community in a research area currently thinks of as being *true*. Each research field has its own "transient gold-standards" which, of course, vary over time. These standards are usually published in reviews, but, due to their time-dependent variability, are not accessible as datasets that can be easily worked with. Instead, this information requires to be extracted from the literature, which gives anyone who tries to work with the current standard a different base-line. This leads to a couple of questions: What is the best way to make this domain-, time-, and data-dependent interaction-information publicly available? Is it possible to find out which interactions are the most relevant? Extracting the relevance of interaction data from a proteome-wide interactome is not a straight-forward task, but a very important challenge as it may assist in finding extended functional clusters in interaction data which may guide the design of experiments.

Here, a first draft of a method is presented that allows the enrichment of a well-known protein-protein interaction network with novel and potentially relevant interactors. Given a protein-protein interaction network, a set of genes that may be related to this network in some way, and an underlying interactome (e. g. parts from a protein-protein interaction database), the method tries to connect the given genes to the network, using interactions from the interaction database, and then reduce this extended network to only the most relevant interactions. In addition, computationally predicted interactions are included as these may introduce edges in the interaction-graph, that have not yet been experimentally reported. An overview of the steps involved in this method is given in Figure 4.1. This generally applicable method was applied to enrich the current view of what the DNA damage response protein-protein interaction network looks like with chromatin-modifying genes from a high-content RNAi-screen. The STRING database was used along with predictions from Scansite 3 to provide additional interaction data. Since there is no dataset of the protein-protein interactions involved in the DNA damage response available, the process of creating this dataset is mentioned here too.

4.1 Terminology

The method presented here uses four different entities of data, each of which will be referred to with different terms. These terms will be defined here:

- The **well-known interaction network** of interest: This is the network that is known (or believed) to be of high confidence and will be referred to as *core-network*, or simply *core-net*.
- **Potentially important genes / interactors**: The genes that are tried to be associated to the core-network are called *target genes*, or short *targets*. The genes in this list are expected to be related to the core-net in some way and should be part of the underlying interaction network (at least partially).

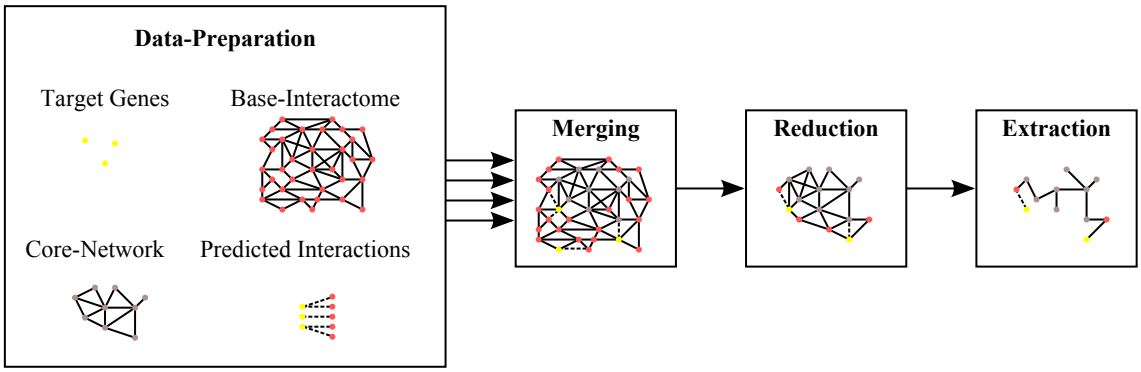


Figure 4.1: Overview of the main steps involved in the method presented in this chapter. After the data was prepared, the different datasets are merged into one network. Predicted interactions that already exist in the underlying network of known interactions are not included in the merged network. In the reduction-step, connections from the target genes to the core-network are searched (here, via a maximum of one intermediate gene). The extraction step then reduces the previously created sub-network to a tree that only contains the most relevant interactions and nodes.

- **Interactome-like network**: The *base-network* (or *base-net*) is a network of known protein-protein interactions. The method expects this network to contain the core-net, or at least overlap with it. Also, this network is expected to be highly connected.
- **Putative interactions**: These *target-interactions* (or *predicted interactions*) are optional to use, but should be included in order to add the possibility to find new interactions. If no predicted interactions are included, the method will just try to associate the target genes with the core-net using base-net interactions. Interactions in this category are either interactions with genes from the core-net or with target genes. This restriction was only introduced to simplify the network.

Although recommendations about the origins of the different groups of data are given it is possible to use any data that can be connected in a way the method can use it. This means, that this method cannot only be applied to protein-protein interaction data, but to basically any data that can be represented in a graph-/network-model and overlaps in the way described above (the base-net should include at least parts of the core-net and target genes). However, the method might not make sense with some kinds of data.

Protein-protein interaction networks are *graphs* that use protein interaction data. For this reason, genes will often be referred to as *nodes* or *vertices*, and interactions between genes as *edges*. Wherever vertex- and edge-scores are used, they will be referred to as v_{type} and e_{type}, with the subscript *type* being the node's or edge's type. Similarly, if an index is given in this annotation ($v_{i;type}$, $e_{i;type}$), only a single value is referred to. The same naming convention is used for other scores (s_{type}).

4.2 Methods

The goal of the method described here is to try to answer the question how a set of genes is connected to a network of proteins that are known to interact closely by associating these with the well-known network and extracting the most relevant interactions. This is achieved in four steps:

1. Prepare network data
2. Associate target genes with core-network (create "subnet")
    - Merge network data
    - Reduce merged network
3. Prepare network-scores
4. Extract relevant interactions from subnet (create tree from graph)

These steps will be described in more detail here, exemplifying the most important points and caveats using the DNA damage response interaction network as core-network and RNAi screen data as targets. The relevance of interactions is determined based on preliminary scores that are associated with the network data in the very beginning. These scores are then prepared for the extraction step after a subnet was created.

4.2.1 Preparing Network Data

The first important step when preparing the network data is to think about how to represent the data. This mainly refers to the representation of nodes (naming convention) in the interaction-graph which is mostly important because node-names should be easily readable and understandable in order to make preparation and visualisation of network data more convenient for the people working with it. Surrogate identifiers (e. g. 0, 1, 2,. . .) or some standardised identifiers — especially those that are primarily numeric — are therefore a bad choice (e. g. ENSG00000149311 and IPI00298306 both represent the ATM kinase) as these will require many additional ID-mapping efforts in every step one wants to take a look at the data. Although the applications that are used here work with any kind of node-identifier, it is important to stick to a single type of IDs in order to avoid ambiguities and hidden duplicates (due to different names for the same nodes). Here, approved gene symbols from the HUGO Nomenclature Commitee[1] (HGNC) are used. These gene symbols are, generally speaking, the gene names used in papers (or very similar to them, such as *CHEK1* instead of *CHK1* for the checkpoint kinase 1) and therefore well-known by researchers in the field. Another thing to consider is that the mapping process itself may result in new duplicates. This issue can be illustrated by the mapping of protein- to gene-identifiers: A single gene may have multiple gene-products, many of which may be known and associated with a separate identifier. A list of protein-identifiers

[1] *http://www.genenames.org/*

can therefore be summarised by the gene that encodes this protein. Mapping a gene-identifier to a protein-identifier is in this case not possible without avoiding ambiguities. However, mapping a protein identifier to a gene symbol is straight-forward. If this mapping process is applied to a network of protein-IDs the mapping to gene symbols inevitably creates duplicate edges wherever multiple nodes are merged into one.

Once a type of identifier was chosen, the different parts of data have to be prepared. This includes the mapping process as well as assigning scores to nodes and edges. The scores that are used can be divided into two different groups with opposite semantics: node-scores, which are the higher the more important the nodes are considered, and edge-scores, which have scores closer to zero the more reliable they are considered. The reason for this use of scores lies in the algorithm that is used in the last step, the extraction of the most relevant interactions. This algorithm creates a Prize Collecting Steiner Tree (PCST) from the enriched core-network by including as many high-score-nodes as possible and avoiding the use of "expensive" edges (i. e. edges with high scores and little evidence). In this problem-formulation, edges are penalised (by their score) if they are included in the network, and nodes are rewarded if they are included. The goal is a tree of maximum weight. This will be described in more detail in Section 4.2.4. Here, node- and edge-scores range approximately from zero to one, but other ranges can be used as well.

Base-Network

The base-net that was used in this application is STRING, restricted to experimentally verified interactions of human proteins. STRING provides a downloadable dataset, but uses Ensembl Protein identifiers (e. g. ENSP00000278616 for ATM). There are a number of reasons why STRING was used instead of any other protein-protein interaction database: First of all, this database summarises data from other PPI databases and therefore excludes the problem of non-overlapping interaction-data across different public databases (as mentioned in Section 2.1.2 on page 16). This is important, because some overlap with the core-net, target genes, and target-interactions is desired. The more nodes the base-net contains, the more likely overlaps are. Secondly, STRING provides downloadable datasets. This is important because the steps described here require locally stored datasets. Third, unlike most other database, the STRING database provides scores for all interactions that give some estimate how likely an interaction is to be true. These scores include information about how many publications an interaction was mentioned in and the experimental method that reported the interaction (e. g. low throughput-methods are more reliable as high-throughput methods), correcting for the probability that an interaction is observed by accident (Von Mering et al., 2005). The availability of interaction scores is crucial for this method as they are used in a later step for determining which interactions are the most relevant. After pre-processing (mapping and removing duplicate genes and interactions) the STRING base-net used here consisted of 12,429 genes and 105,990 interactions.

Preliminary Scores. The downloadable dataset provides scores (s_{STRING}) between 0 and 1000, with higher scores meaning higher confidence. Here, these values are reversed and mapped to $[0;1]$:

$$e_{base} = 1 - \frac{s_{STRING}}{1000}$$

These scores give information about the interaction-confidence and are used only to describe interactions. All node-scores in the base-network are set to zero ($v_{base} = 0$), as these should neither be favoured nor disfavoured in the PCST-step.

Core-Network

A core-net can either be a self-defined network or a precompiled interaction-list. In any case, the core-network should be thought of correct, excluding interactions that are not thought of as important. Here, the current knowledge of the DNA damage response was intended to be used. Unfortunately, no dataset describing this network is available in the public domain. Parts of the network (i. e. certain pathways) are available in the Kyoto Encyclopædia of Genes and Genomes[2] (KEGG) and some review-papers show different parts of the interaction network in different levels of detail, but no complete or even close-to-complete network was found. For this reason, a network was manually created by extracting binary protein-protein interactions mentioned in different reviews and putting them together in a network. Based on the assumption that interactions and genes mentioned in reviews from the past couple of years in high impact factor journals present the state-of-the-art view of the field, reviews like these were used as starting points for the network. A list of key references is shown in Table 4.1 However, extracting an interaction network from reviews is not a straight-forward task. Missing references, opposing statements in different publications, fuzzy wording, and ambiguous phrasing are only a few of the reasons that make the manual curation process quite hard and time-consuming. Additional ambiguities are added due to different gene names and naming conventions. These problems required the use of additional resources (another 40 publications).

In order to be able to track interactions back to their origin, each interaction (as pair of interacting genes) was annotated with a list of references. In addition, the type of interaction (e. g. phosphorylation) was stored along with the interaction, wherever this information was given. Edge directions and temporal aspects were ignored. Hence, the edges in the network only represent interactions, no matter when they happen during the DNA damage response, and what role they play. It can be concluded that nodes of a higher degree (number of outgoing edges) are more important. However, this is a dangerous assumption as the presence of more edges does not necessarily mean that there are no other genes that have a higher degree or are more important, but just that this gene (and its direct interactors) has been studied in detail since it is currently thought of as important.

Preliminary Scores. Scores in the core-network are determined in a later step, meaning that the initial values defined for this network do not matter.

[2] *http://www.kegg.jp/*

Table 4.1: A list of references that served as starting points in creating a high-confidence protein-protein interaction network of the DNA damage response

Key References in Creating the DNA Damage Response Core-Network

Title	Reference
The Smc complexes in DNA damage response	Wu and Yu (2012)
Susceptibility pathways in Fanconi's anemia and breast cancer	D'Andrea (2010)
The DNA damage response: making it safe to play with knives	Ciccia and Elledge (2010)
Kinases that control the cell cycle in response to DNA damage: Chk1, Chk2, and MK2	Reinhardt and Yaffe (2009)
ATR: an essential regulator of genome integrity	Cimprich and Cortez (2008)
The DNA damage response: ten years after	Harper and Elledge (2007)
The role of double-strand break repair — insights from human genetics	O'Driscoll and Jeggo (2006)
Molecular mechanisms of mammalian DNA repair and the DNA damage checkpoints	Sancar et al. (2004)
Regulating mammalian checkpoints through Cdc25 inactivation	Donzelli and Draetta (2003)
ATM and related protein kinases: safeguarding genome integrity	Shiloh (2003)

Target Genes

The genes used in this work as targets are derived from a high-throughput, high-content quantitative automated microscopy assay which was applied to an RNAi-screen of chromatin-modifying enzymes and interacting proteins (Floyd et al., 2012). Changes in chromatin structure are known to be important events in the response to DNA damage. However, the molecular details of how chromatin is altered in response to DNA single- and double strand breaks are not known in detail yet. In their work, the authors used 44 384 well plates (with knockdowns for 2209 genes), each well containing on average 550 cells, resulting in about 9.3 million ($44 \cdot 384 \cdot 550$) cells. DNA damage was induced in these cells using ionising radiation and the effects were examined at four time-points (no irradiation, and 1, 6, and 24 hours after irradiation) with measurements of 615 features that covered five phenotypic readouts: DNA damage induced signalling, cell cycle status, mitotic entry, apoptosis, and mitotic progression. In the analysis of that enormous amount of data (2 TB of image data in a single experiment), first the most significant features are chosen, followed by a hit identification method that extracts those genes that showed the most statistically significant *jRiger* false discovery rate (Rameseder, 2012). The jRiger-method is based on a gene set enrichment analysis described in Subramanian et al. (2005). The target genes used here are the top 110 hits found by this method.

Preliminary Scores. The values provided by the jRiger-method $s_{jRiger} \subset \mathbb{N}^+$ are used to calculate scores for the target genes:

$$v_{i;target} = \frac{s_{i;jRiger}}{\max(s_{jRiger})}$$

Target Interactions

In order to allow this method to introduce potentially important edges that have not yet been identified in experiments and which may be important missing links between pathways or functional units, predicted interactions are added to the base-network. Basically any prediction tool can be used to create a list of predicted interactions. Here, the Scansite 3 framework was used for this purpose. The reasons are obvious: First of all, Scansite focuses on kinase-substrate interactions, which are — as emphasised many times before — incredibly important in signal transduction networks. Secondly, scores are needed that allow to compare the predicted results to each other and to other interactions in the base-net. And, of course, the prediction tool should be computationally accessible, as predictions have to be repeated for many proteins. The availability of the Scansite 3-framework was thus also an important point. Two strategies have to be considered when deciding where predicted interactions should be included: interactions with target genes and interactions with core genes (both these categories may, of course, overlap). If the core-network is expanded, a huge number of interactions will be included (given the size of the core-network) which will unnecessarily blow up the network's size. Adding predicted interactions to target genes, however, adds more possible starting points to connect the targets to the core-net. Given the assumption that all target genes are of interest, but nothing is known about which core-nodes are the most important the latter strategy makes more sense, which is also why this strategy is used and described here. Although easily possible, the combination of both methods has not been tried yet. Using Scansite, a high-stringency *Protein Scan* has been performed for every target gene. All the interactions predicted this way were then included in the data set of predicted interactions. For this purpose a Java application that uses the Scansite 3 framework was created and used. One protein can contain a motif several times, with different scores at each site. Here, only the best score is saved and used.

Preliminary Scores. After a protein was scored with all mammalian motifs, the scores from this search ($s_{Scansite}$) are mapped to the range $[0;1]$.

$$e^I_{i;predicted} = \frac{s_{i;Scansite}}{\max(s_{Scansite}) - \min(s_{Scansite})}$$

This is done in order to be able to compare all protein scan searches to each other. Since high Scansite scores do not have the node-score semantic that is required in this method (good Scansite scores are close to zero, whereas good node-scores are closer to one), the scores are reversed for all vertices:

$$v_{predicted} = 1 - e^I_{predicted}$$

.

4.2.2 Associating Target Genes with Core-Network

Once all datasets are prepared, the targets can be associated with the core-network. This is done using interactions from the base-network and predicted interactions and can be divided into two steps: (1) Merging all the data together into one large network, and (2) extracting a sub-network (also referred to as extended core-network, or simply "subnet") from this network. This is done by a Java application (*SubNetter* application) that was created for this very purpose.

The merging step starts with reading the base-network. If edges occur multiple times, the mean of their scores is used. This applies both to genes and interactions. The same is done for all other datasets. Once all datasets are in memory, they are merged. In the merging process, scores are overwritten, and the nodes' type information is set. This is done to be able to easily visualise the network in a later step (which only makes sense if the network is not too large). First, the core-net is added to the base-net. In this step the core-network has priority and overwrites all overlapping nodes' scores in the base-network. Then, the target genes are marked in (or added to) the base-net, again having priority. Lastly, the target interactions are merged, not overwriting any node- or edge-information, only adding *new* edges and nodes.

Once all the networks are merged into one, the subnetting-step is started. The algorithm that does this works like this: Starting from every target gene in the network, a path along all edges to all of this node's neighbours and all of their neighbours is "walked" recursively. If the core-network is reached using less than a given number of intermediate nodes, the nodes on this path are added to the subnet. In order to make sure that all the most important parts are in the resulting sub-network, the core-network and the target genes are added to the subnet in the very beginning. In this step, two important parameters come into play: *maximum path length* and *maximum edge score*. The maximum path length defines how many steps a target is allowed to be away from the core-network (number of intermediate nodes minus one). For example, if a node X is connected to Y via two nodes A and B (X–A–B–Y) and the maximum path length is 2, the nodes are not added to the subnet, whereas if this parameter has a value of 3, the nodes on this path are added to the sub-network. The other parameter defines a score-limit for edges that are used on a path like this. If, for example, a value of 0.5 is given as maximum edge score, only edges with a score less or equal to 0.5 are considered in the path. Edges with a higher score are ignored.

In the end of this step, the core-net is, dependent on the parameter-settings, associated with all, none, or a subset of the target genes. This network is the starting point for the last step, the extraction of the potentially most relevant interactions.

4.2.3 Preparing Network Scores

Before the SubNetter application writes the subnet to the filesystem in different file-formats it prepares the edge- and vertex-scores for the next step. This is necessary, because the scores used in the network come from different sources and thus different distributions. An evaluation of what scores are the most important needs to be done at this point. Starting with vertex-scores, there are

three different vertices in the subnet: nodes from the core-network, target genes, and intermediate genes (either from the base-net or from target-interactions). The most important vertex-scores are those of the target genes, since these are the ones this method focuses on. These, along with the core-network nodes, are the ones the end-result should contain. Intermediate nodes are not specifically favoured (most of all, because no prior knowledge about their importance is available), but are welcome to show up in the result, especially because they are the links between the other two node-categories. Hence, it makes sense to make all node-scores dependent on the targets' scores. Similarly, there are three types of edges in the subnet: intra-core interactions, predicted interactions, and other out-of-core interactions. Here, the most relevant values are the out-of-core edge-scores, since they are derived from the base-network and thus come from the STRING database. Edges within the core should be slightly favoured as they are known to be important. Predicted interactions should be favoured in a similar manner, as these offer the inclusion of novel interactions.

As mentioned previously, vertex-scores can be interpreted as profits, and edge-scores as costs. Therefore, a score of 0 means the same for nodes and edges, i. e. that they are neither favoured nor disfavoured. A high score means a strong favouring for nodes to be included in the final result, but a disfavouring for edges. By combining these ideas into simple formulas the following vertex-scores were used:

$$v_{i;target} = \frac{s_{i;jRiger}}{\max(s_{jRiger})} \qquad \text{(as defined before)}$$

$$v_{intermediate} = 0$$

$$v_{core} = median(v_{target}) + MAD(v_{target})$$

Interaction-scores are calculated in a similar manner:

$$e_{intermediate} = e_{base}$$

$$e_{core} = median(e_{intermediate}) - 2 \cdot MAD(e_{intermediate})$$

$$e_{i;predicted} = \frac{e^I_{i;predicted} - \min(e^I_{predicted})}{\max(e^I_{predicted}) - \min(e^I_{predicted})} \cdot e_{core}$$

The scoring convention used here is just a simple approach to combine these ideas, but is far from perfect. In this first version of the method, however, the focus was not to produce the best possible results, but to show that the method works and is worth following-up on.

4.2.4 Extraction of Relevant Interactions by creating a Prize Collecting Steiner Tree

At this point, a protein-protein interaction network is available that contains the core-network with additional interactors, including the desired target genes of interest. Starting from this, the question that is asked is, how can this network be reduced to only the most relevant interactions? Given the size of the network created in the preceding step (thousands of genes and interactions), it is necessary to have a computational tool that assists in identifying important genes and interactions

from the network. The relevance of interactions is determined by the scores given in the network. Here, the network was reduced by building a Prize Collecting Steiner Tree (PCST).

The PCST problem is a special version of the Minimum Steiner Tree problem which is defined as follows: Given a connected, undirected graph $G = (V, E, c)$ that consists of a set of edges E, vertices V, a cost function that defines edge-distances $c : E \mapsto \mathbb{R}^{\geq 0}$ and a set of Steiner nodes $S \subseteq V$, a Steiner Tree $T_S = (V_S, E_S)$ which is a subgraph of G is searched with $S \subseteq V_S \subseteq V$ and $E_S \subseteq \{(v_1, v_2) \mid (v_1, v_2) \in E, \{v_1, v_2\} \subseteq V_S\}$ (Mehlhorn, 1988). In other words, a *Steiner Tree* is a tree subgraph of G that contains all nodes S. Paths in G are defined as sequences of vertices $v_1, v_2, \ldots, v_n$ of V with $\forall_{1 \leq i < n} i : (v_i, v_{i+1}) \in E$. Lengths of paths are the sum of the distances of the edges along the path. A graph's total distance is the sum of all distances. G_S is called a minimal Steiner Tree if G_S is a Steiner Tree and the tree's total distance is minimal. Finding a minimal Steiner Tree has been shown to be an NP-complete problem (Garey and Johnson, 1979).

The primary target when building a minimal Steiner Tree is to have *all* Steiner nodes connected to each other. This goal is changed in the more specialised formulation of the *Prize Collecting Steiner Tree* problem: Given a graph $G = (V, E, p, c)$ with V, E, and c as above and p being a profit-function $p : V \mapsto \mathbb{R}^{\geq 0}$, a Prize Collecting Steiner Tree $T_P = (V_P, E_P)$ is searched that does not minimise the graph's total distance, but instead maximises the graph's profit function:

$$profit(T_P) = \sum_{v \in V_P} p(v) - \sum_{e \in E_P} c(e)$$

The most important difference in this version of the problem is that there is no set of Steiner vertices S involved. Instead, the profits of the nodes determine whether a node will be included in the tree. In the application of this problem here, the target genes are considered to be Steiner vertices, which do not necessarily have to be included in T_P. Those that are included in the T_P are referred to as S_P ($S_P \subseteq S \subseteq V_P$), those that are not included as $\bar{S_P} = S \setminus S_P$.

The PCST problem can be applied to any network structure and became popular in the context of telecommunication companies wanting to find a cost-efficient way to expand their network in a city: In this example, vertices V are blocks of buildings and edges E are the connecting infrastructure that has to be built by the company. Hence, the profits p (associated with buildings) are the charges that the company expects the people using their infrastructure to pay, the costs c (associated with edges) are the costs that are associated with connecting the buildings. It is in the company's interest to only include blocks of buildings that they can expect a high profit from by at the same time avoiding to build unnecessary connections. The original graph G in this example includes all the various ways to connect all the buildings in an area.

A number of methods have been developed that create a Minimum Steiner Tree from a set of genes in a surrounding protein-protein interaction network (White and Ma'ayan, 2007; Berger et al., 2007; Chen et al., 2012). However, using these methods it is not possible to find out which genes are of relevance with respect to a known network, since all targets are connected in this approach. This is why the method described here uses an application that solves the PCST problem.

Four different applications that involve a solver for the PCST problem were tried; one of these

applications, DHEA (Ljubic et al., 2005, 2006), is an application that implements an optimal solver for this problem and is used by two of the other applications that were tried. However, since this application uses commercial software libraries for some optimisation steps, this application could not be used. The next application that was tried was the *heinz package*[3] ("heavy induced subgraphs") which uses the DHEA software in a Python library to solve the maximum-weight connected subgraph problem (Dittrich et al., 2008). Here, again, the missing software libraries posed a problem since the application requires to be run locally. The same group extended this Python library to an R-package (*BioNet*), replacing the heuristics from the commercial software libraries with algorithms available in R (Beisser et al., 2010). Unfortunately, the method provided in this package did not allow the use of self-defined edge-scores which was the reason why another method was searched for. SteinerNet[4] (Huang and Fraenkel, 2009), a web-application that uses the DHEA-program to build a PCST from a given scored network is the application that was used in the end. The SteinerNet-output was then visualised and analysed using Cytoscape[5].

4.3 Results

One of the key results of this project is the DNA damage response PPI network that was manually created. Figure 4.2 shows a Cytoscape-view (Shannon et al., 2003) of the network. This view not only gives an overview of the most important genes that are involved in the response to DNA damage, but also allows to quickly identify which of these important gene products certain proteins interact with. The interaction-data can easily be annotated with interaction types and other additional information and thus is not just a useful resource for the work described here, but also for everyone else who works with DNA damage response data. It is easy to identify those genes that are studied the most, which indicates that they are very important in this context. These genes are highly interconnected *hubs* in the network: *ATM* kinase, one of the most important players in the DNA damage response, can easily be found in the centre of the figure. Other important hub-like genes are *ATR*, *BRCA1/2*, *H2AFX*, *TP53*, *PARP1/2*, *PRKDC* and *CHEK1/2*, most of which have been mentioned in the overview of the DNA damage response in Section 2.2.

In addition, some functional complexes and clusters can be identified easily by just looking at the network. This includes, for example, the *CDK2*/*CDC25* cluster in the bottom right corner where many genes involved in the initiation and maintenance of the G1/S checkpoint arrest (Sancar et al., 2004) are displayed. Another functional complex that is easy to see without much effort is the *PARP1/2* cluster which has been primarily studied because of its important role in the sensing of DNA single strand breaks. PARP-inhibitors are used as cancer therapeutics (for example, for breast and ovarian cancers). They block the single strand break repair pathway (because the necessary sensors are turned off) in cancerous cells, which may result in lethal (for the cell) double strand breaks. In healthy cells the PARP-repair pathway is backed up by another repair pathway (e. g. NHEJ or HR). Another big important cluster can be seen in the bottom left corner: The cluster of

[3] *http://www.mi.fu-berlin.de/w/LiSA/Heinz*
[4] *http://fraenkel.mit.edu/steinernet/*
[5] *http://cytoscape.org/*

FANC-genes shows genes involved in the Fanconi's anemia (FA) repair pathway. FA has become an attractive model for studying breast cancer susceptibility genes because of similarities of FA genes to *BRCA*-genes (Wang, 2007). Two protein complexes that are also worth mentioning can also be found easily: The MRN-complex, consisting of *MRE11A*, *RAD50*, and *NBN* — located inside the "arc" on the top right side — is one of the first molecular reactions to DNA double strand breaks. The 911-complex has a similar function at single strand breaks and consists of *RAD9A*, *HUS1*, and *RAD1*. This complex can be found in the centre of the lower third of the figure.

When describing the *SubNetting*-step, two parameters were mentioned: The maximum path length and the maximum edge score. These two options offer a way to change the size of the resulting network. The greater the value for the maximum path length is chosen, the more nodes are allowed to be used as intermediates between core-net and targets, which, due to the high interconnectedness of protein-protein interaction networks, not only results in a greater number of proteins and edges included in the subnet, but also in a longer runtime. At the same time, a high number of intermediate nodes increases the probability of finding a connection for every node. A way to restrict the number of edges and nodes included in the network is to filter by the score of the edges that are used. Table 4.2 shows these results based on runs with maximum path lengths of 2 and 3, meaning one and two intermediates, respectively. Also, the scores were either not restricted at all, or with a value of 0.3.

Table 4.2: Starting from a merged network (merged base, core, targets, and target-interactions) with 12,483 vertices ($| V |$) and 106,627 ($| E |$) the SubNetter-application created networks of the sizes listed in the table. In addition, the runtime for the application to do so, and the number of excluded target genes ($| S | = 111$) are listed. The abbreviations PLR and ESR stand for Path Length Restriction and Edge Score Restriction, respectively. The given runtime values are the rounded mean-values of 20 runs.

SubNetting-Results for Different Settings

PLR / ESR	$\| V_P \|$	$\| E_P \|$	$\| \bar{S_P} \|$	**Runtime**
2 / none	1,137	14,957	22	18 s
2 / 0.3	605	3,902	25	3 s
3 / none	3,247	43,664	21	1,460 s
3 / 0.3	1,414	9,628	24	97 s

Surprisingly, the number of target nodes that the application was not able to connect to the core-network ($| \bar{S_P} |$) did not vary too much across the different settings. Based on how the SubNetting-method works it is not surprising however, that the nodes that were excluded overlapped entirely in the settings listed in the table. That is, the run with the most exclusions excluded all the nodes that were also excluded in the run with the next smaller number of exclusions, and so on. It is safe to assume that these nodes are indeed not too closely connected to the DNA damage response PPI network. The application's runtime mostly depends on the combination of parameters that is chosen. Generally speaking, a high maximum path length requires a very strongly restricted edge-score. If the quality of the sub-network is considered based on the number of target genes that were not connected, it is easy to see that even with settings that vary just between two values the

network sizes and runtimes change a lot, with always including about three quarters of the target genes. This is why the most restrictive and fastest setting was chosen to create the subnet that was used in the next and final step.

The SteinerNet web application reduced this network to the tree shown in Figure 4.3. Most nodes in the tree are core-nodes, which is not surprising, since these nodes are (1) the bigger part of the subnet that was used as input for SteinerNet and (2) the scores were chosen in a way to favour them. The most important nodes to look at in this network are the target nodes, and the way they are connected to the core-nodes. 28 targets (from $\mid S_P \mid = 86$ in the subnet) stayed in the network. Some of these nodes were found to directly interact with core or base genes, others were connected via predicted interactions with core or base genes. The way the subnet was reduced shows a severe separation between core genes and target genes (top/bottom in the figure) which can be interpreted in at least two ways: To start with, there may be functional differences in the two sets of nodes. The core focuses on DNA damage response, the target genes are mostly genes that are known to be involved in chromatin modification. This may be one reason why there is a separation in the tree. Alternatively, the scores may have been chosen poorly. As mentioned before, the scores are calculated in an easy way to test the method in this first instance and, since the scores are what the PCST step is based on, the results may look completely different if the scores are changed. At least those parts of the tree that are connected via experimentally verified edges can be considered true, including connections via nodes from the base-net (*INPP5E*, *CTU1*, *PI4K2A*, *PIK3R4*, *LRGUK*, *ZGPAT*, *SCYL3*, *PKIB*, *CSNK1G3*, and *SPRPR*). Genes that are connected to the core-net (directly or via base-nodes) with predicted interactions may be true as well. *UCK1* and *CLK3* are especially important results, since they are potentially relevant links to the DNA damage response network that have not been verified yet (or not reported in the dataset used in this work). All the interactions and genes below (in the figure) the *SRPR–AURKB* arm have to be handled carefully as they are mainly connected via predicted interactions and do not include any core genes at all.

In any case, the question whether the genes that were connected to the core-net using this method are actually involved in the DNA damage response needs different methods to be answered. At this point of the work presented here it is too soon to draw conclusions about the biological significance of the tree that was created. Before this can be done, more effort needs to be spent on scoring techniques and more parameter-combinations for creating sub-network have to analysed with respect to the sub-networks that are created.

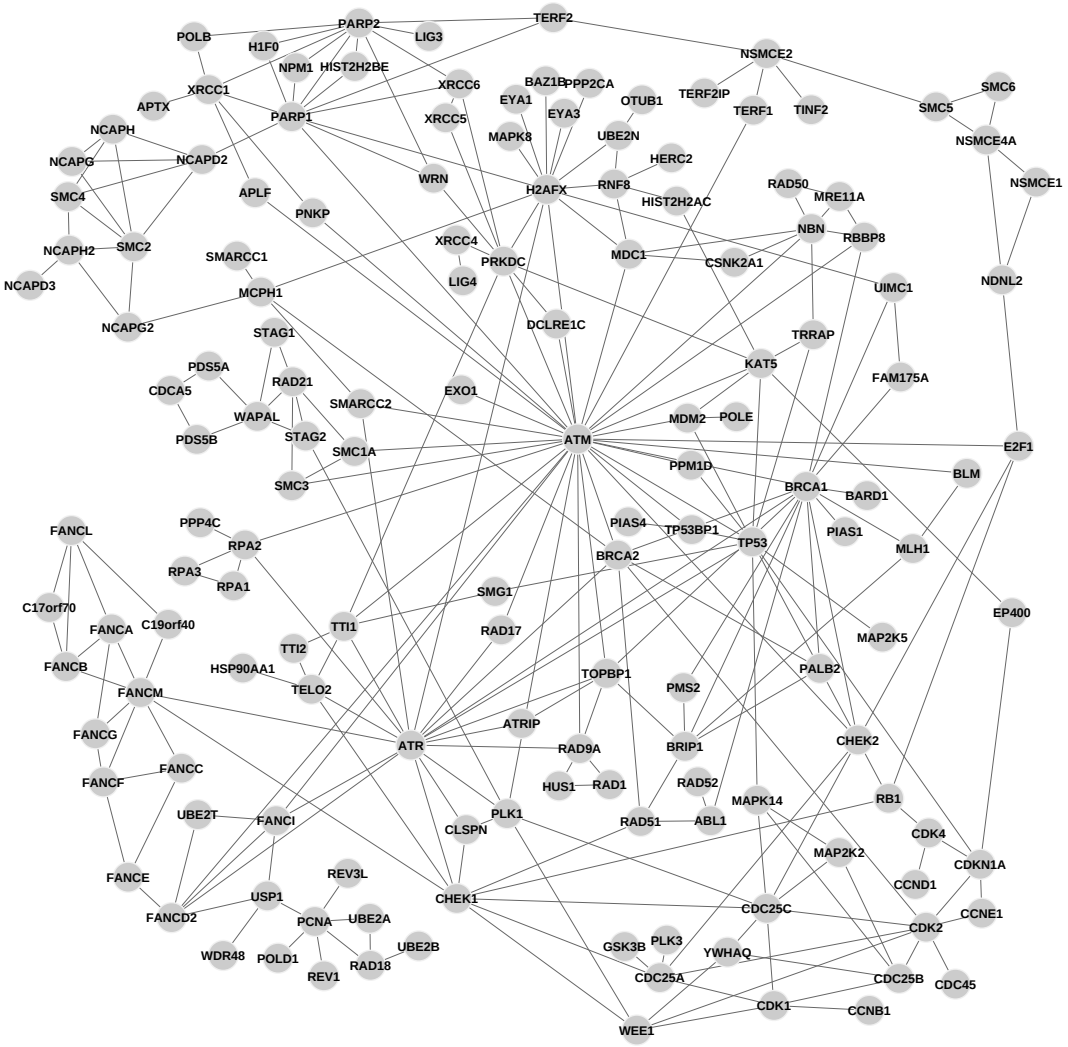


Figure 4.2: The manually created DNA damage response core-network. Edges in this view only mean that in some publication (noted in the original binary interaction data set) evidence was found that the connected genes interact. There is no temporal or directional component given in that graph.

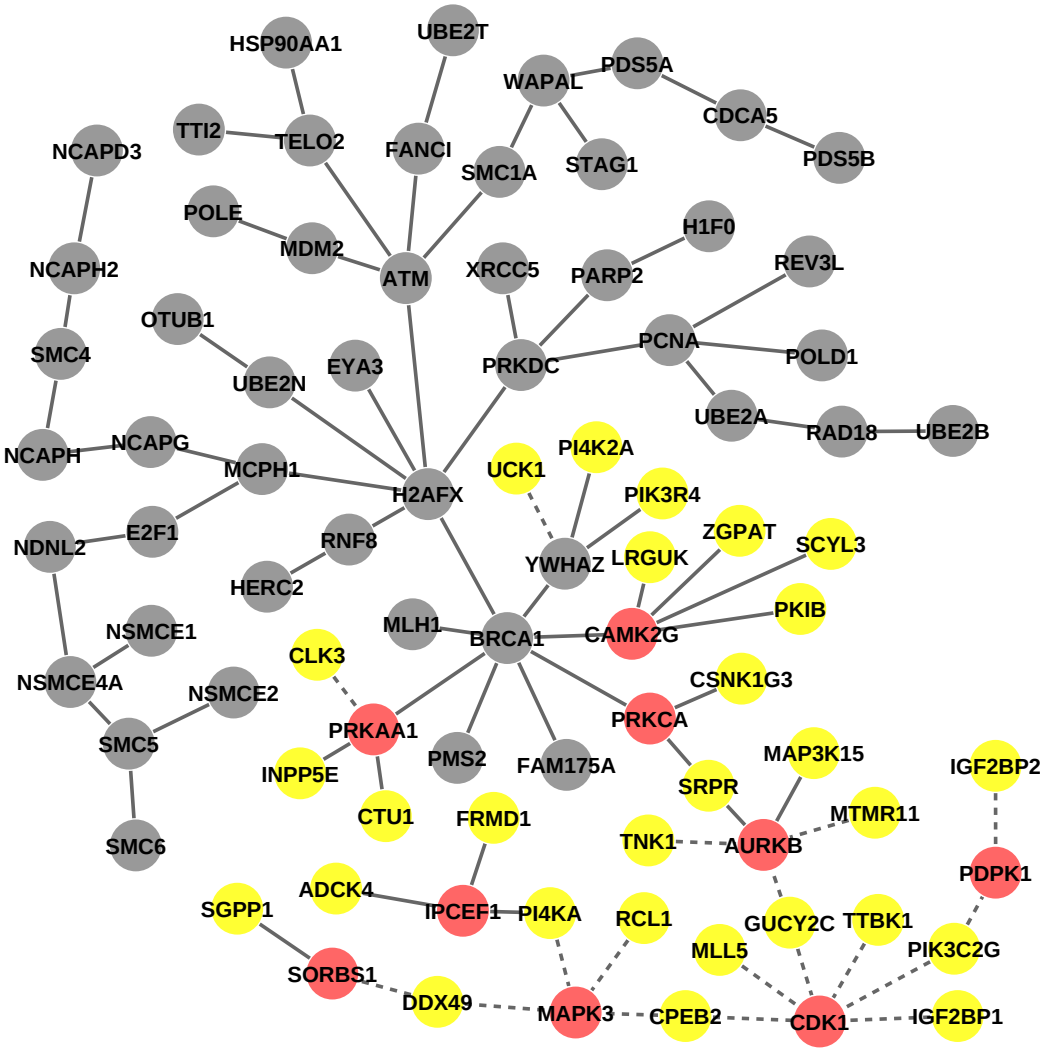


Figure 4.3: The tree created by SteinerNet. Nodes from the core-network are displayed in grey, nodes from the base-net in light red, and targets in yellow. Interactions that were added by Scansite are indicated by a dashed line between nodes. This figure was created with Cytoscape.

Chapter 5

Conclusion

The preceding chapters introduced protein-protein interactions in general and their relevance in biological networks exemplified on the DNA damage response. One specific interaction-type, phosphorylation, was described in more detail. Phosphorylation is one of the most important post-translational modification in cell signalling networks as it provides molecular switches that, combined in a network, allow a very fine-grained control of what happens in a cell. However, identifying phosphorylations experimentally is not always easily possible and takes a lot of effort, both in time and costs. Hence, tools that allow a high-confidence prediction of phosphorylation are necessary to help scientists design experiments and target their research in an efficient and reliable way. This work focused on one tool like that: Scansite 3 is based on the identification of short linear sequence-motifs recognised by kinases and binding domains, a concept that has been proven to be useful and valid since the first version of Scansite was made available online in 2001. In further consequence, the Java-based Scansite-suite that was created in this work and presented here was used to associate a set of genes derived from an RNAi-screen with a manually created DNA damage response protein-protein interaction network via known and putative interactions. Although this method needs some more refinement and tweaking in its application, it is a first part of an algorithmic pipeline that will be created for putting genes (e. g. outputs from knockdowns or other experiments) in their place in a network the researchers are interested in. This method has potential to allow scientists to quickly identify which genes (out of potentially thousands) are most relevant to the model (represented by a network of protein-protein interactions) they are studying, which allows to save time and costs, and helps to decide which results are worth following up on ("low-hanging fruit strategy").

The kinase-substrate interaction prediction tool Scansite 3 is now available online for researchers all over the world. Users are encouraged to report bugs, and submit ideas about how Scansite can be improved. Since the entire Scansite suite is implemented in Java, it is easy to introduce new features without worrying about the interplay between different technologies. A number of new features are described in Chapter 3 on page 30, all of which assist users in identifying sites and help them to decide whether these sites are *real*. Still, it is important to always bear the caveats of the method applied in mind: Scansite's predictions are based on the assumption that peptide library data is

sufficient information for predicting an interaction, and that the experimentally derived PSSM-data is correct. Hence, Scansite's predictions, and results from computational prediction tools in general, have to be scrutinised with extreme caution before they are included in the design of (potentially expensive and time-consuming) experiments. Structural information, as well as other factors, is important for interactions to happen *in vivo*. This is why Scansite provides a variety of supporting information for the sites it predicts, including a site-score and -percentile, domain information, a surface-accessibility value, general information about motifs and proteins, and links to previously mapped sites. Links to other bioinformatics tools allow users to quickly "get a second opinion" about sites in the protein they are analysing. Different kinds of additional supporting information were thought about in detail, and will be available in future versions of Scansite:

- **Subcellular localisation** information can be added to sites if the kinase and putative substrate are found in the same cellular compartments. Obviously, only proteins that physically meet in a cell are able to interact. This idea can be extended in a way that also co-expression is considered.
- Similarly, protein **interaction networks** can be used to identify which substrates are phosphorylated by the same kinases, which allows to draw conclusions about how likely potential interactions are to be true.
- Another important factor when looking at phosphorylation sites is **evolutionary conservation**. This is taken into account in Scansite only in a very subtle way (BLAST-search for site sequences). Algorithms are available for determining if a single amino acid is conserved (e. g. Zhang et al., 2007), but no model has been published so far that extends this idea to an entire motif. A model like this may improve Scansite's predictions even further.
- Including parts of this data or the data that Scansite already provides in Scansite's **scoring function** (and providing this score in addition to the site-only score) would make it easier to keep track of this whole arsenal of different information.

These are just a few ideas that were considered but needed to be deferred due to time-restrictions on the work presented here. Scansite is one of the most widely used tools for identifying short linear sequence-motifs in proteins and has been the "quasi-gold standard" in almost every publication that compared different kinds of kinase-substrate interaction prediction tools in the past years.

The application of Scansite 3 in the enrichment of already known interaction networks with new genes showed that Scansite's data can be used in different ways. Even by just using the web-interface, it is possible to build potentially relevant signalling cascades / pathways by incrementally repeating *Protein Scans* and *Database Searches*. This work described a new way to use Scansite's data by embedding predicted interactions in a network view. The goal in this application was to create a first step in an algorithmic pipeline that allows researchers to associate genes with an interaction network in a quick and easy way. The method has been shown to work, but it still needs some improvements: To start with, the scoring-strategy for nodes and edges needs to be improved in a way that is more statistically comprehensible, meaning that the system chosen at the moment is

based only on a simple idea that suits the PCST-scoring scheme, but does not take the distributions of the initial scores (as they come from different sources) into consideration. More sophisticated scores will probably yield better and more biologically relevant results. Secondly, the parameters of the SubNetter application require some experimenting: The maximum path length, of course, is highly dependent on how close the genes are expected to be related with the network that is looked at. Finding a good score-threshold, however, is a much harder task as it is highly dependent on the scores that are used and where the data is taken from. Also, different parameters to control the SubNetting-step could be introduced. For example, a limit on node-scores, a limit on the score of a path (which is the sum of the edges in the path), or a limit on the number of publications that have reported certain edges. The method could also be greatly improved, if more data was considered in general. At the moment, only binary protein-protein interaction data and associated scores are looked at. But this data can be enriched with GO-terms, colocalisation data, gene-expression data, or other biologically relevant information. Including more data, however, comes at the cost of increased complexity, which could make the method less comprehensible.

Once the method has been finalised, a detailed analysis of the links shown in the resulting PCST is required to check if the results are indeed biologically relevant. This ultimate question is what drove the development of this method in the first place. When analysing the tree-data, one may consider also taking a look at the subnet that was used to create the tree (if it is not too large). This may help understanding the tree on a different level and gives an overview of which genes are the most important hubs in that intermediate network. However, it is important to mention that this method cannot only be applied to biological data, but to any kind of network data that is available. In a far-fetched example, this method could be applied to user-interaction data in social networks in order to answer the following question in a hiring process of companies: Based on who users (targets: applicants for job) in a social network (base-network) interact with, how likely is it that people fit in this new social environment (core-net: the company, and social interactions within)? Of course, scores have to be calculated in a completely different way in other applications, but the idea persists.

Finally, one of the most important results in this work was the construction of a high-confidence network that represents the current view of what genes are important in the molecular response to DNA damage and how these genes interact with one another. Since nothing like this has been published before, it is a valuable resource for anyone who works with DNA damage response data, and provides a helpful overview for people who are new to this very field. It can be used as a quick-reference for biologists, or as a dataset for computational analyses. This network needs to be kept up-to-date as new publications in that research area will show that new interactions are important and others may be proven to be wrong (or insignificant).

In conclusion, this work described a novel way to associate genes to a network of interest and gave an overview of the new version of Scansite which can now be accessed online. The network-enrichment project will be followed up on by Jonathan Rameseder in his thesis project (Rameseder, 2012).

References

Alexander, J., Lim, D., Joughin, B., Hegemann, B., Hutchins, J., Ehrenberger, T., Ivins, F., Sessa, F., Hudecz, O., Nigg, E., Fry, A., Musacchio, A., Stukenberg, P., Mechtler, K., Peters, J., Smerdon, S., and Yaffe, M. (2011). Spatial exclusivity combined with positive and negative selection of phosphorylation motifs is the basis for context-dependent mitotic signaling. *Sci Signal. 2011 Jun 28;4(179):ra42.*

Altschul, S., Gish, W., Miller, W., Myers, E., and Lipman, D. (1990). Basic Local Alignment Search Tool. *J Mol Biol. 1990 Oct 5;215(3):403-10.*

Bairoch, A. (1992). PROSITE: a dictionary of sites and patterns in proteins. *Nucleic Acids Res. 1992 May 11;20 Suppl:2013-8.*

Beisser, D., Klau, G., Dandekar, T., Müller, T., and Dittrich, M. (2010). Bionet: an r-package for the functional analysis of biological networks. *Bioinformatics. 2010 Apr 15;26(8):1129-30. Epub 2010 Feb 25.*

Berger, S., Posner, J., and Ma'ayan, A. (2007). Genes2Networks: connecting lists of gene symbols using mammalian protein interactions databases. *BMC Bioinformatics. 2007 Oct 4;8:372.*

Bjellqvist, B., Hughes, G., Pasquali, C., Paquet, N., Ravier, F., Sanchez, J., Frutiger, S., and Hochstrasser, D. (1993). The focusing positions of polypeptides in immobilized pH gradients can be predicted from their amino acid sequences. *Electrophoresis. 1993 Oct;14(10):1023-31.*

Blom, N., Sicheritz-Pontén, T., Gupta, R., Gammeltoft, S., and Brunak, S. (2004). Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. *Proteomics. 2004 Jun;4(6):1633-49.*

Burnett, G. and Kennedy, E. (1954). The enzymatic phosphorylation of proteins. *J Biol Chem. 1954 Dec;211(2):969-80.*

Chen, E., Xu, H., Gordonov, S., Lim, M., and Ma'ayan, M. P. A. (2012). Expression2Kinases: mRNA profiling linked to multiple upstream regulatory layers. *Bioinformatics. 2012 Jan 1;28(1):105-11. Epub 2011 Nov 10.*

Cherry, J., Hong, E., Amundsen, C., Balakrishnan, R., Binkley, G., Chan, E., Christie, K., Costanzo, M., Dwight, S., Engel, S., Fisk, D., Hirschman, J., Hitz, B., Karra, K., Krieger, C., Miyasato, S., Nash, R., Park, J., Skrzypek, M., Simison, M., Weng, S., and Wong, E. (2011). Saccharomyces

Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res. 2012 Jan;40(Database issue):D700-5. Epub 2011 Nov 21.*

Ciccia, A. and Elledge, S. (2010). The DNA damage response: making it safe to play with knives. *Mol Cell. 2010 Oct 22;40(2):179-204.*

Cimprich, K. and Cortez, D. (2008). ATR: an essential regulator of genome integrity. *Nat Rev Mol Cell Biol. 2008 Aug;9(8):616-27. Epub 2008 Jul 2.*

D'Andrea, A. (2010). Susceptibility pathways in Fanconi's anemia and breast cancer. *N Engl J Med. 2010 May 20;362(20):1909-19.*

De las Rivas, J. and Fontanillo, C. (2010). Protein-protein interactions essentials: key concepts to building and analyzing interactome networks. *PLoS Comput Biol. 2010 Jun 24;6(6):e1000807.*

Dinkel, H., Chica, C., Via, A., Gould, C., Jensen, L., Gibson, T., and Diella, F. (2011). Phospho.ELM: a database of phosphorylation sites — update 2011. *Nucleic Acids Res. 2011 Jan;39(Database issue):D261-7. Epub 2010 Nov 9.*

Dittrich, M., Klau, G., Rosenwald, A., Dandekar, T., and Müller, T. (2008). Identifying functional modules in protein-protein interaction networks: an integrated exact approach. *Bioinformatics. 2008 Jul 1;24(13):i223-31.*

Donzelli, M. and Draetta, G. (2003). Regulating mammalian checkpoints through Cdc25 inactivation. *EMBO Rep. 2003 Jul;4(7):671-7.*

Fields, S. and Song, O. (1989). A novel genetic system to detect protein-protein interactions. *Nature. 1989 Jul 20;340(6230):245-6.*

Floyd, S., Pacold, M., Clarke, S., Blake, E., Fydrych, A., Ho, R., Lee, M., Root, D., Carpenter, A., Sabatini, D., Chen, C. F. J. B. C., and Yaffe, M. (2012). The bromodomain protein Brd4 insulates chromatin from DNA damage signaling through acetyl-lysine binding. *Manuscript in Revision*.

Garey, M. and Johnson, D. (1979). *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman and Company.

Gnad, F., Ren, S., Cox, J., Olsen, J., Macek, B., Oroshi, M., and Mann, M. (2007). PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. *Genome Biol. 2007;8(11):R250.*

Harper, J. and Elledge, S. (2007). The DNA damage response: ten years after. *Mol Cell. 2007 Dec 14;28(5):739-45.*

Hermjakob, H. (2006). The HUPO proteomics standards initiative–overcoming the fragmentation of proteomics data. *Proteomics. 2006 Sep;6 Suppl 2:34-8.*

Hoeijmakers, J. (2009). DNA damage, aging, and cancer. *N Engl J Med. 2009 Oct 8;361(15):1475-85.*

Hornbeck, P., Kornhauser, J., Tkachev, S., Zhang, B., Skrzypek, E., Murray, B., Latham, V., and Sullivan, M. (2012). PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res. 2012 Jan;40(Database issue):D261-70. Epub 2011 Dec 1.*

Hu, Z., Narayanaswamy, M., Ravikumar, K., Vijay-Shanker, K., and Wu, C. (2005). Literature mining and database annotation of protein phosphorylation using a rule-based system. *Bioinformatics. 2005 Jun 1;21(11):2759-65. Epub 2005 Apr 6.*

Huang, S. and Fraenkel, E. (2009). Integrating proteomic, transcriptional, and interactome data reveals hidden components of signaling and regulatory networks. *Sci Signal. 2009 Jul 28;2(81):ra40.*

Hunter, S., Apweiler, R., Attwood, T., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L., Finn, R., Gough, J., Haft, D., Hulo, N., Kahn, D., Kelly, E., Laugraud, A., Letunic, I., Lonsdale, D., Lopez, R., Madera, M., Maslen, J., McAnulla, C., McDowall, J., Mistry, J., Mitchell, A., Mulder, N., Natale, D., Orengo, C., Quinn, A., Selengut, J., Sigrist, C., Thimma, M., Thomas, P., Valentin, F., Wilson, D., Wu, C., and Yeats, C. (2009). InterPro: the integrative protein signature database. *Nucleic Acids Res. 2009 Jan;37(Database issue):D211-5. Epub 2008 Oct 21.*

Hutti, J., Jarrell, E., Chang, J., Abbott, D., Storz, P., Toker, A., Cantley, L., and Turk, B. (2004). A rapid method for determining protein kinase phosphorylation specificity. *Nat Methods. 2004 Oct;1(1):27-9.*

Iakoucheva, L., Radivojac, P., Brown, C., O'Connor, T., Sikes, J., Obradovic, Z., and Dunker, A. (2004). The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res. 2004 Feb 11;32(3):1037-49. Print 2004.*

Kerrien, S., Aranda, B., Breuza, L., Bridge, A., Broackes-Carter, F., Chen, C., Duesbury, M., Dumousseau, M., Feuermann, M., Hinz, U., Jandrasits, C., Jimenez, R., Khadake, J., Mahadevan, U., Masson, P., Pedruzzi, I., Pfeiffenberger, E., Porras, P., Raghunath, A., Roechert, B., Orchard, S., and Hermjakob, H. (2012). The IntAct molecular interaction database in 2012. *Nucleic Acids Res. 2012 Jan;40(Database issue):D841-6. Epub 2011 Nov 24.*

Koh, G., Porras, P., Aranda, B., Hermjakob, H., and Orchard, S. (2011). Analyzing protein-protein interaction networks. *J Proteome Res. 2012 Apr 6;11(4):2014-31. Epub 2012 Mar 2.*

Lai, A., Ba, A. N., and Moses, A. (2012). Predicting kinase substrates using conservation of local motif density. *Bioinformatics. 2012 Apr 1;28(7):962-9. Epub 2012 Feb 1.*

Lee, J. and Paull, T. (2007). Activation and regulation of ATM kinase activity in response to DNA double-strand breaks. *Oncogene. 2007 Dec 10;26(56):7741-8.*

Levy, E., Landry, C., and Michnick, S. (2010). Cell signaling. signaling through cooperation. *Science. 2010 May 21;328(5981):983-4.*, (20489011).

Licata, L., Briganti, L., Peluso, D., Perfetto, L., Iannuccelli, M., Galeota, E., Sacco, F., Palma, A.,

Nardozza, A., Santonico, E., Castagnoli, L., and Cesareni, G. (2012). MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res. 2012 Jan;40(Database issue):D857-61. Epub 2011 Nov 16.*

Lindahl, T. (1993). Instability and decay of the primary structure of DNA. *Nature. 1993 Apr 22;362(6422):709-15.*

Linding, R., Jensen, L., Ostheimer, G., van Vugt, M., Jørgensen, C., Miron, I., Diella, F., Colwill, K., Taylor, L., Elder, K., Metalnikov, P., Nguyen, V., Pasculescu, A., Jin, J., Park, J., Samson, L., Woodgett, J., Russell, R., Bork, P., Yaffe, M., and Pawson, T. (2007). Systematic discovery of in vivo phosphorylation networks. *Cell. 2007 Jun 29;129(7):1415-26. Epub 2007 Jun 14.*

Ljubic, I., Weiskircher, R., Pferschy, U., Klau, G., Mutzel, P., and Fischetti, M. (2005). Solving the prize-collecting Steiner tree problem to optimality. *Proceedings of the Seventh Workshop on Algorithm Engineering and Experiments (ALENEX 05).*

Ljubic, I., Weiskircher, R., Pferschy, U., Klau, G., Mutzel, P., and Fischetti, M. (2006). An algorithmic framework for the exact solution of the prize-collecting Steiner tree problem. *Mathematical Programming*, 105(2):427–449.

Mathivanan, S., Periaswamy, B., Gandhi, T., Kandasamy, K., Suresh, S., Mohmood, R., Ramachandra, Y., and Pandey, A. (2006). An evaluation of human protein-protein interaction data in the public domain. *BMC Bioinformatics. 2006 Dec 18;7 Suppl 5:S19.*

Mehlhorn, K. (1988). A faster approximation algorithm for the Steiner problem in graphs. *Information Processing Letters 1988; 27(3);125 - 128.*

Mok, J., Kim, P., Lam, H., Piccirillo, S., Zhou, X., Jeschke, G., Sheridan, D., Parker, S., Desai, V., Jwa, M., Cameroni, E., Niu, H., Good, M., Remenyi, A., Ma, J., Sheu, Y., Sassi, H., Sopko, R., Chan, C., Virgilio, C. D., Hollingsworth, N., Lim, W., Stern, D., Stillman, B., Andrews, B., Gerstein, M., Snyder, M., and Turk, B. (2010). Deciphering protein kinase specificity through large-scale analysis of yeast phosphorylation site motifs. *Sci Signal. 2010 Feb 16;3(109):ra12.*

Obenauer, J., Cantley, L., and Yaffe, M. (2003). Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res. 2003 Jul 1;31(13):3635-41.*

O'Driscoll, M. and Jeggo, P. (2006). The role of double-strand break repair - insights from human genetics. *Nat Rev Genet. 2006 Jan;7(1):45-54.*

Pflieger, D., Gonnet, F., de la Fuente van Bentem, S., Hirt, H., and de la Fuente, A. (2010). Linking the proteins — elucidation of proteome-scale networks using mass spectrometry. *Mass Spectrom Rev. 2011 Mar-Apr;30(2):268-97. doi: 10.1002/mas.20278. Epub 2010 May 24.*

Prasad, T., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., Balakrishnan, L., Marimuthu, A., Banerjee, S., Somanathan, D., Sebastian, A., Rani, S., Ray, S., Kishore, C. H., Kanth, S., Ahmed, M., Kashyap, M., Mohmood, R., Ramachandra, Y., Krishna, V., Rahiman, B., Mohan, S., Ranganathan, P.,

Ramabadran, S., Chaerkady, R., and Pandey, A. (2009). Human Protein Reference Database — 2009 update. *Nucleic Acids Res. 2009 Jan;37(Database issue):D767-72. Epub 2008 Nov 6.*

Punta, M., Coggill, P., Eberhardt, R., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E., Eddy, S., Bateman, A., and Finn, R. (2012). The Pfam protein families database. *Nucleic Acids Res. 2012 Jan;40(Database issue):D290-301. Epub 2011 Nov 29.*

Rameseder, J. (2012). PhD-Thesis in Preparation. *Massachusetts Institute of Technology.*

Reinhardt, H. and Yaffe, M. (2009). Kinases that control the cell cycle in response to DNA damage: Chk1, Chk2, and MK2. *Curr Opin Cell Biol. 2009 Apr;21(2):245-55. Epub 2009 Feb 21.*

Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M., and Séraphin, B. (1999). A generic protein purification method for protein complex characterization and proteome exploration. *Nat Biotechnol. 1999 Oct;17(10):1030-2.*

Rousseeuw, P. and Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, pages 1273–1283.

Salwinski, L., Miller, C., Smith, A., Pettit, F., Bowie, J., and Eisenberg, D. (2004). The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res. 2004 Jan 1;32(Database issue):D449-51.*

Sancar, A., Lindsey-Boltz, L., Unsal-Kaçmaz, K., and Linn, S. (2004). Molecular mechanisms of mammalian DNA repair and the DNA damage checkpoints. *Annu Rev Biochem. 2004;73:39-85.*

Schneider, T. and Stephens, R. (1990). Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res. 1990 Oct 25;18(20):6097-100.*

Shannon, P., Markiel, A., Ozier, O., Baliga, N., Wang, J., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res. 2003 Nov;13(11):2498-504.*

Shiloh, Y. (2003). ATM and related protein kinases: safeguarding genome integrity. *Nat Rev Cancer. 2003 Mar;3(3):155-68.*

Songyang, Z., Blechner, S., Hoagland, N., Hoekstra, M., and Cantley, H. P.-W. L. (1994). Use of an oriented peptide library to determine the optimal substrates of protein kinases. *Curr Biol. 1994 Nov 1;4(11):973-82.*

Stark, C., Breitkreutz, B., Chatr-Aryamontri, A., Boucher, L., Oughtred, R., Livstone, M., Nixon, J., van Auken, K., Wang, X., Shi, X., Reguly, T., Rust, J., Winter, A., Dolinski, K., and Tyers, M. (2011). The BioGRID Interaction Database: 2011 update. *Nucleic Acids Res. 2011 Jan;39(Database issue):D698-704. Epub 2010 Nov 11.*

Stelzer, G., Dalah, I., Stein, T., Satanower, Y., Rosen, N., Nativ, N., Oz-Levi, D., Olender, T., Belinky, F., Bahir, I., Krug, H., Perco, P., Mayer, B., Kolker, E., Safran, M., and Lancet, D.

(2011). In-silico human genomics with GeneCards. *Hum Genomics. 2011 Oct;5(6):709-17.*

Subramanian, A., Tamayo, P., Mootha, V., Mukherjee, S., Ebert, B., Gillette, M., Paulovich, A., Pomeroy, S., Golub, T., Lander, E., and Mesirov, J. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A. 2005 Oct 25;102(43):15545-50. Epub 2005 Sep 30.*

Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., Doerks, T., Stark, M., Muller, J., Bork, P., Jensen, L., and von Mering, C. (2011). The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res. 2011 Jan;39(Database issue):D561-8. Epub 2010 Nov 2.*

Trost, B. and Kusalik, A. (2011). Computational prediction of eukaryotic phosphorylation sites. *Bioinformatics. 2011 Nov 1;27(21):2927-35. Epub 2011 Sep 16.*

Turinsky, A., Razick, S., Turner, B., Donaldson, I., and Wodak, S. (2010). Literature curation of protein interactions: measuring agreement across major public databases. *Database (Oxford). 2010 Dec 22;2010:baq026. Print 2010.*

Uversky, V. and Dunker, A. (2010). Understanding protein non-folding. *Biochim Biophys Acta. 2010 Jun;1804(6):1231-64. Epub 2010 Feb 1.*

Von Mering, C., Jensen, L., Snel, B., Hooper, S., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M., and Bork, P. (2005). STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res. 2005 Jan 1;33(Database issue):D433-7.*

Wang, W. (2007). Emergence of a dna-damage response network consisting of fanconi anaemia and brca proteins. *Nat Rev Genet. 2007 Oct;8(10):735-48. Epub 2007 Sep 4.*

Ward, I. and Chen, J. (2004). Early events in the DNA damage response. *Curr Top Dev Biol. 2004;63:1-35.*

Weterings, E. and Chen, D. (2008). The endless tale of non-homologous end-joining. *Cell Res. 2008 Jan;18(1):114-24.*

White, A. and Ma'ayan, A. (2007). Connecting seed lists of mammalian proteins using steiner trees. *Signals, Systems and Computers, 2007. ACSSC 2007. Conference Records 155-159, 4-7 Nov. 2007.*

Wu, N. and Yu, H. (2012). The Smc complexes in DNA damage response. *Cell Biosci. 2012 Feb 27;2(1):5.*

Yaffe, M., Leparc, G., Lai, J., Obata, T., Volinia, S., and Cantley, L. (2001). A motif-based profile scanning approach for genome-wide prediction of signaling pathways. *Nat Biotechnol. 2001 Apr;19(4):348-53.*

Zhang, S., Pan, Q., Cheng, Y., Zhang, Y., and Chou, K. (2007). An improved algorithm for estimation of residue evolutionary conservation. *Bioinformatics and Biomedical Engineering. ICBBE 2007.*, pages 80–83.