Published in final edited form as:
Cell Biochem Biophys. 2009 ; 53(2): 101–114. doi:10.1007/s12013-009-9042-y.

In-silico construction of a protein interaction landscape for nucleotide excision repair

Nancy Tran[1], **Ping-Ping Qu**[2], **Dennis A. Simpson**[1], **Laura Lindsey-Boltz**[3], **Xiaojun Guan**[4], **Charles P. Schmitt**[4], **Joseph G. Ibrahim**[2,5,6], and **William K. Kaufmann**[1,5,6]

[1]*Department of Pathology and Laboratory Medicine, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina.*

[2]*Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina.*

[3]*Department of Biochemistry and Biophysics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina.*

[4]*Renaissance Computing Institute, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina.*

[5]*Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina.*

[6]*Center for Environmental Health and Susceptibility, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina.*

Abstract

To obtain a systems-level perspective on the topological and functional relationships among proteins contributing to nucleotide excision repair (NER) in *Saccharomyces cerevisiae*, we built two models to analyse protein-protein physical interactions. A recursive computational model based on set theory systematically computed overlaps among protein interaction neighborhoods. A statistical model scored computation results to detect significant overlaps which exposed protein modules and hubs concurrently. We used these protein entities to guide the construction of a multi-resolution landscape which showed relationships among NER, transcription, DNA replication, chromatin remodeling, and cell cycle regulation. Literature curation was used to support the biological significance of identified modules and hubs. The NER landscape revealed a hierarchical topology and a recurrent pattern of kernel modules coupling a variety of proteins in structures that provide diverse functions. Our analysis offers a computational framework that can be applied to construct landscapes for other biological processes.

Introduction

Mounting effective defences to environmental challenges and the repair of DNA lesions are critical cellular functions required to maintain genome stability for normal cell growth [1]. DNA repair involves a broad spectrum of cellular mechanisms affecting signaling pathways of various biological processes such as DNA replication, cell cycle regulation, transcription [2], and chromatin remodelling [3]. Nucleotide excision repair (NER) is a cellular mechanism that removes a wide variety of DNA lesions including the major UV-induced DNA photo-products. Failure of NER in xeroderma pigmentosum is associated with over 1,000-fold increased incidence rate of skin cancer [4]. Much progress has been made in identifying various components and interactions in the NER machinery 5. However, integrated views of these

Correspondence should be addressed to W.K.K. (wkarlk@med.unc.edu)..

interactions at multiple levels of resolution and from a systems biology perspective are needed for understanding the contributions of various biological processes in NER. Such views can provide new insights into unknown protein functions and pathways involving NER and cellular DNA damage response. It can also guide researchers to potential targets for drug therapy to inactivate NER.

The wealth of protein interactions available in public databases provides a tremendous opportunity, but also poses a challenge in constructing these system-level views. For example, visualizing 1,000 interactions among 100 proteins (assuming 10 interactions/protein on the average) often results in a fuzzy ball that is difficult to decipher topologically and functionally. When a protein interacts with many other proteins that participate in diverse functions (e.g., based on the *Saccharomyces* Genome Database [6] (SGD, http://www.yeastgenome.org), the master cell cycle regulator CDC28 has been found to interact with ≈250 proteins), determining the subset of proteins that have biological relevance to a cellular mechanism of interest (e.g. NER) is challenging.

To build a topological and functional landscape for NER, we created a set-theory-based computational model and a statistical model to systematically analyse NER protein-protein physical interactions in *Saccharomyces cerevisiae.* One approach to analyzing protein interaction networks uses graph theory to study graphs, which are mathematical structures that are used to model pairwise relations between objects from a certain collection. Another approach that was used in this report is set theory, the branch of mathematics that studies sets, which are collections of objects. The set-theory-based statistical model exploited protein interaction overlaps to uncover protein modules (sets of proteins in which each protein interacts with every other protein) and hubs (proteins with many interactions engaging multiple biological processes). A kernel module was defined as a fully-connected set of proteins that can couple with multiple proteins independently to achieve functional variance. These protein entities were used to identify topological and functional relationships among yeast proteins retrieved from SGD. SGD integrates protein-protein physical interactions curated from various small-scale experiments and high-throughput studies that used diverse identification techniques such as yeast two-hybrid, co-immunoprecipitation, and mass spectrometry.

Methods

To select NER-related genes in yeast, 34 human NER genes were obtained from a DNA repair gene list [7] and used to search via BLAST [8] for yeast homologs against all verified open reading frames in SGD. Because there can be multiple homologs for a given human NER protein, 121 yeast homologs (p-value $<10^{-12}$) were found. Inclusion of multiple homologs offered the potential of disclosing yeast proteins with possible roles in NER that otherwise might be missed. In addition, one functional homolog (TFB5, 7) and ten other yeast proteins annotated with NER functions in SGD were added, producing a total of 132 yeast NER-related proteins, 38 of which are essential. Among the remaining 94 non-essential proteins, 20 were associated with UV-sensitivity when depleted [6,9]. These results are provided in Table 1.

A recursive set-based model for computing protein interaction overlaps

Our analysis was built upon the premise that, if two proteins share many physical interaction partners, it is likely that they also share similar functions [10] or that they operate in the same pathway(s). Based on this premise, we constructed a set-theory based model using set intersections to systematically find proteins that are shared among NER protein neighborhoods. For this purpose, we defined the direct neighborhood *NA* of a protein *A* as the set of all proteins that directly interact with *A*, including *A* itself (Fig. 1a). Using this definition, interaction partners common to proteins *A*, *B*, and *C* can be computed as the intersection set $N_A \cap N_B \cap N_C$ of three neighborhoods.

We applied recursion to find all possible intersections among 131 neighborhoods of the yeast NER-related proteins (MRK1 was the only protein in the group of 132 homologs that did not have physical interactions recorded in SGD as of Jan. 2007). Recursion allowed reductions in computation costs by re-using results of the previous, non-empty intersections for the next round of computation, into which additional neighborhoods were systematically incorporated. The sizes of intersections became smaller as the number of neighborhoods increased. Computation stopped when all intersections were empty or all neighborhoods had been incorporated, whichever came first.

Denoting k as the number of neighborhoods participating in the computation of intersections, the number of distinct intersections is the number of distinct combinations resulting from choosing k out of the 131 neighborhoods, i.e.,

$$\binom{131}{k} = \frac{131!}{k!\,(131 - k!)} \quad (k=2, \cdots 131)\,.$$

Figure 1b illustrates an example with four protein neighborhoods. Because proteins are selective in their binding and specific in their interactions with molecular targets [11], many combinations produced empty intersections, making the proposed recursive method computationally feasible.

A statistical model for scoring and detecting significant overlaps

To detect statistically significant neighborhood overlaps/intersections that were produced by the computational model, we built a statistical model to score overlap results and perform tests seeking evidence against the null hypothesis of no overlap among protein neighborhoods. Towards this goal, let X_k be a random variable representing the number of proteins shared among k neighborhoods ($N_1 \ldots N_k$). Because any protein in these neighborhoods was either included in the intersection or excluded - a binary outcome - X_k can be modeled via a binomial distribution. To estimate the probability of protein sharing, our model took into account neighborhood sizes and the sizes of associated non-empty intersections. As a result, the probability of observing at least s shared proteins in an intersection of k neighborhoods is:

$prob\,(X_k \geq s) = \sum_{i=s}^{m} \binom{m}{i} {p_k}^i (1 - p_k)^{m-i}$ where $m=min(|N_1,\cdots|N_k|)$. $s > 0$, $k \geq 2$ $\mathrm{p_k}$ = prob. of sharing a protein among k neighborhoods

Because an intersection cannot have more proteins than the smallest of the k participating neighborhoods, m must be the minimum size of the k neighborhoods (Fig. 1a, a similar concept of minimum neighborhood size has been successfully used in analyzing metabolic networks [10]). Hence, $\hat{p_k}$ can be estimated as the average proportion of shared proteins relative to the smallest of the k neighborhoods, for all combinations of choosing k out of 131 neighborhoods. Denoting t as the total number of non-empty intersections, s_i as the number of shared proteins observed for a particular intersection i, and m_i as the size of the associated smallest neighborhood, we have: $\widehat{p_k} = \frac{\sum_t^{i=1} S_i/m_i}{\binom{131}{k}}$ (Derivation of this equation is explained in Supplementary Equation 1).

To control the expected proportion of incorrect rejections among all the rejections - false discovery rate (FDR) - made during statistical tests, we applied the Benjamini-Hochberg (BH) procedure [12,13] on the p-values ($prob(X_k \geq s)$) computed for neighborhood intersections, assuming the null hypothesis is true. When testing hundreds or thousands of times or more, FDR is a less stringent criterion than FWER (family wise error rate) in including more test data that otherwise might be missed [13].

To apply the BH procedure, p-values associated with a given *k*-neighborhood were first sorted in ascending order. Then the BH equation:

$$\text{cutoff}\quad\text{threshold}\quad \widehat{l}=\max_{1\le l\le t}\left(l\!:p-value\le\alpha\times\frac{l}{t}\right)$$

was evaluated based on three parameters , a pre-chosen α level, the location (l) of a p-value in the sorted list, and the total number (t) of p-values associated with non-empty intersections (Table 2, column 2). For this application, because the probabilities of protein sharing among neighborhoods were small (Table 2, column 3), it is reasonable to choose a small α of 0.001× % = 10−5 to control the false discovery rate. Selection of small α 's has been successfully used to increase the robustness of protein complex identification [14].

If $\hat{l}$ existed, the null hypothesis was rejected for all the p-values in the sorted list that were less than or equal to the p-value associated with $\hat{l}$ (i.e., p-value$_1$ ≤ p-value$_2$ ≤ … p-value $\hat{l}$); otherwise, no rejection was made. The BH procedure was repeated for all the p-values for a given k (k = 2 … 15, shown in Table 2, column 1). We defined two scoring functions: protein sharing scores as $-\log_{10}(\ p$ – values) and BH cutoff scores as $-log_{10}\left(\alpha\times\frac{l}{t}\right)$ to facilitate comparisons between small p-values and parameter values in the BH equation.

Finally, statistically significant results associated with rejected p-values were further reduced when neighborhood intersections from smaller *k* values were subsets of those from larger *k*'s. For example, if a protein set {*V,W*} was found to be the intersections of 3 neighborhoods ($N_A \cup N_B \cup N_C$) and subsequently of 4 neighborhoods ($N_A \cup N_B \cup N_C \cup N_C$), the 3-neighborhood result was redundant and discarded.

Supplementary Table 1 provides a list of significant and non-redundant overlaps, along with their protein sharing scores and BH cutoff scores. These results were obtained from a software prototype that we implemented for the models described above, using the C++ object-oriented standard template library (STL) [15]. STL provides container templates that support set-based objects and operations on these objects, including intersection and membership. 131 set objects, each representing a neighborhood, were dynamically selected from ≈35,000 physical interactions and created in memory. The number of neighborhoods producing non-empty intersections ranged from 2 to 15 (Table 2, column 1) and the number of non-empty intersections varied from 1 to ≈17,500 (Table 2, column 2).

Uncovering protein modules and hubs using significant overlap results

Modules can be uncovered by exploiting the relationships among proteins in significant neighborhood intersections and associated core proteins (Supplementary Table 1). Defined as a fully-connected set in which every protein interacts with all other proteins in the set, a module can be identified using the simple rules below:

> A neighborhood intersection contains a subset of associated core proteins. For example, let *A*, *B*, *C* be core proteins and the intersection $N_A \cup N_B \cup N_C$ of associated neighborhoods be *{ A,B,C,V,W }*. *A*, *B*, and *C* form a module because *A*, being in the intersection, interacts with all core proteins, i.e., *B* and *C*; similarly, *B* interacts with *C*.

Rule 1 holds and the intersection has one or more additional proteins/modules relative to the core-protein set or vice versa. Continuing with the above example, $\{A,B,C,V\}$ and $\{A,B,C,W\}$ constitute two larger modules because V and W are in the intersection, hence they interact with core proteins A, B, and C. Because V and W share the same module, it is likely that they also interact (Fig. 2). This prediction can be readily verified by querying the membership of W in the neighborhood N_v or vice versa (i.e., $W \in N_V | V \in N_W$). Without the V-W interaction, $\{A,B,C,V,W\}$ would still form a densely connected module instead of one that is fully-connected.

If an identified protein module is a subset of a larger module, the latter supersedes the former, allowing construction of larger aggregates. Such aggregation reduces redundancy and is useful to identify modules with many subunits. Supplementary Table 2 provides a list of modules identified using the rules given above.

Hubs can be loosely defined as connectors linking proteins in adjacent neighborhoods to provide and coordinate diverse functions. Significant neighborhood intersections and associated core protein sets that contain a single protein/module interacting with many proteins are hub candidates. Towards building multi-resolution views of interactions among proteins contributing to NER, we focused on those candidates that interacted with, or were components of, identified modules. A candidate that was a module component was selected if it interacted with many other proteins in addition to those in its own module. An example is the RAD14 hub which interacts with RAD1, RAD10, RAD16, RAD3, RFA1, TFB1, in addition to proteins in the module [RAD14, RAD23, RAD4, RAD7].

Not only single proteins can be hubs, modules or members of aggregates that satisfy the criteria given above can also be hubs. For example, the [RFC2-5] module interacts with several components of DNA replication (CTF18, CTF8, ECO1, ELG1, POL30), DNA damage checkpoint (RAD24), and a component of chromatin remodeling (ASF1). A list of hubs along with details on hub selections are provided in Supplementary Table 3.

Guided by the identified modules and hubs, and using literature curation to support their biological significance, we built a multi-resolution landscape for NER, visualized via Cytoscape [16]. We focused on the topological and functional relationships among protein modules and hubs within the NER and transcription system, and the relationships of these protein entities with other biological processes, based on gene ontology annotations in SGD [6] (Supplementary Tables 2, 3).

Results

Relationships among modules and hubs within the NER and transcription system

Modules and hubs identified for NER and transcription were hierarchically organized. At the top of the hierarchy was the [RAD14, RAD3, MSH2] module (magenta triangle in Fig. 3a) which linked to three other main module groups, DNA damage sensors, the TFIIH complex, and helicases/nucleases.

RAD14 is a hub connecting various DNA damage sensor modules—RAD14 and the RAD23-RAD4 complex independently recognize DNA damage and recruit other NER proteins to DNA lesions [5]. RAD23, possessing a ubiquitin-like domain that interacts with the proteasome and two sequences that bind to ubiquitin; it is thought to deliver ubiquitinated substrates to the proteasome [17]. When cells are damaged, RAD23 inhibits the ubiquitin-mediated degradation of RAD4, resulting in RAD4's stabilization [18]. Several TFIIH subunits (SSL2, TFB1-2) also interact with RAD23.

The [RAD14, RAD7, RAD16] module leads to a sensor subnetwork that includes ELC1 and ABF1. In the global genome NER pathway, the [RAD7, RAD16, ELC1] module participates in DNA damage recognition and the regulation of RAD23′s ubiquitination [19]. The module was suggested to function as both an ATPase and an E3 ligase, congruent with the observations that RAD7 is similar to an F-Box protein, and RAD16 has a RING domain [19]. In the transcription-coupled NER pathway, the ELC1 elongation factor promotes the ubiquitination and degradation of RNA POL 2 blocked at damage sites [20]. On the other hand, the [RAD7, RAD16, ABF1] module mediates NER [21], transcription, and chromatin remodeling via ABF1 [22]. ABF1 controls nucleosome positioning, keeping regions of chromatin in non-transcribed DNA free of nucleosomes to facilitate repair [22].

RAD3 is a hub connecting all known subunits of TFIIH except TFB5—TFIIH is a ten-subunit transcription factor with seven basal subunits (TFB1,2,4,5, RAD3, SSL1, SSL2) and a trimer complex (TFB3-CCL1-KIN28) that phosphorylates the C-terminal domain of RNA POL2 [23]. Figure 3b shows the interactions of TFIIH subunits with RAD3, which is a 5′ to 3′ helicase/ATPase involved in DNA unwinding to facilitate NER and allows transcription initiation proteins to access DNA. SSL2 encodes a 3′ to 5′ counterpart [24]. TFIIH subunits are organized around the [RAD3, TFB3, TFB4, CCL1, SSL1] module, which acts as a kernel to which TFB1 and KIN28 associate independently, producing two fully-connected aggregates of six subunits. This pattern led to the definition of a kernel module as a fully-connected set of proteins that can couple with multiple macromolecules independently to achieve functional variance. TFB1 provides an alternate path to NER sensors RAD14 and RAD23, supplementing the main path via RAD3. TFB1 also forms a separate module with TFB2, TFB4, and TFB5. On the other hand, KIN28 interacts with RPO21 (the largest subunit of RNA polymerase 2) which, in turn, interacts with NER sensors RAD14, RAD23, and RFA1.

Besides its global genome NER functions, RAD3 was also found in the transcription-coupled NER pathway. It interacts with MET18 and RAD26. MET18 regulates TFIIH to influence NER and transcription [25]. RAD26 is a SWI2/SNF2 ATPase and is needed during transcription for the displacement/removal of stalled RNA polymerases to allow repair proteins to access DNA [26].

MSH2 is a hub sharing helicases/ATPases and nucleases with NER—Two modules relate MSH2 to NER, [RAD14, RAD1, RAD10, MSH2] and [RAD3, RAD2, SSL2, MSH2]. They indicate that MSH2 connects with all known helicases and nucleases involved in NER, although MSH2 was originally found to be required for mismatch repair [27]. Through association with the RAD14 sensor, the endonuclease RAD1-RAD10 is targeted to DNA damage sites [28].

Relationships of NER-transcription modules and hubs with other biological processes

Protein modules and hubs identified for NER and transcription also interact with many proteins involved in three major biological processes - DNA replication, chromatin structure and remodeling, and cell cycle regulation - and one miscellaneous category. Figure 4 provides a high-level view of these relationships. More detailed views are expanded in Figure 5. Functions in the miscellaneous category (the last column of Fig. 4b) include ubiquitin-dependent protein catabolism (RPN6), actin cytoskeleton organization (VPS1), nuclear import of cargo proteins (KAP95), mRNA export from nucleus (YRA1), and GTP biosynthesis (IMD3).

Relationships with DNA replication—HPR5, MSH2, POL30, SMT3 (Fig. 4a) along with RFA1-3 are multi-function junction hubs bridging NER and transcription to DNA replication. From these hubs emerge four main functional groups (Fig. 5a and Supplementary Table 2), the RFA group, the DNA polymerase group, the POL30 group and the RFC group.

RFA1 and RFA2 form a kernel that interacts with the DNA damage checkpoint proteins LCD1 and MEC1, and with MSH2, MSH6 and DNA2. The MSH2-MSH6 complex senses and corrects DNA replication errors [27] caused by mis-incorporation or by misalignment/slippage. DNA2, a DNA replication factor with ATPase, nuclease, and helicase activities [6], is thought to be involved in DNA double-strand break and post-replication repair [29].

The DNA polymerase group is required for DNA synthesis and includes subunits of POLδ ([CDC2, HYS2, POL32] module), POLε ([DPB2-4, POL2] module), and POLσ (TRF5).

POL30/PCNA confers processivity to DNA polymerases δ and ε during DNA replication [30]. Along with these polymerases, POL30 participates in DNA synthesis for NER after excision of DNA lesions [31]. On the other hand, in the absence of DNA damage, sumo-modified PCNA (SMT3 is a protein of the SUMO family [6]) preferentially binds the HPR5 helicase. This binding disrupts RAD51 nucleoprotein filaments to inhibit the RAD52-dependent recombinational pathway, thereby preventing undesired recombination of replicating chromosomes in normal cells [32]. Interactions between PCNA and RAD27/FEN1 stimulate FEN1's nuclease activity on flap substrates in the presence of RFC and ATP [33].

RFC loads the sliding clamp PCNA onto DNA. Similar to RFA1-2, the RFC2-5 subunits form a kernel module to which associate RFC1 and other proteins - RAD24, ECO1, the CTF8-CTF18-DCC1 complex, ASF1, and ELG1. Three of these proteins replace RFC1 to produce alternative RFC complexes. Specifically, the [RAD24, RFC2-5] module is a complex that loads the 9-1-1 sliding clamp at sites of damage to mediate DNA damage checkpoints [34]. The [CTF8, CTF18, RFC2-5] module participates in the establishment of sister chromatid cohesion [34]. In the [ELG1, RFC2-5] module, ELG1 functions redundantly with RAD24 in response to DNA damage and in activating the checkpoint kinase RAD53 during S phase [35]. Finally, RFC recruits the nucleosome assembly/disassembly factor ASF1 to DNA, and together they affect the completion of DNA synthesis upon DNA damage [36].

Relationships with chromatin remodeling—The functions of chromatin remodeling in NER and transcription (Fig. 5b) are organized around the casein kinase II holoenzyme. Subunits of this kinase (CKA1-2, CKB1-2) couple with various histones and other proteins, linking chromatin remodeling to NER via ABF1, to transcription regulation via SPT15 and CHD1, and to DNA replication via ASF1.

SPT15 is a TATA binding protein and a component of the RNA POL 1 core factor, of TFIID and TFIIIB, all of which are required for transcription by RNA POL 1, 2, and 3 respectively, as summarized in SGD. Identified as a hub, SPT15 is in the intersection of six NER-transcription related protein neighborhoods - CKA2, MOT1, STH1, RAD23, TFB2, TFB4 (Supplementary Table 1). MOT1 is a SWI2/SNF2 ATPase and a transcription regulator that displaces SPT15 from DNA [37]. STH1 is an ATPase component of the RSC chromatin remodeling complex with functions in transcription regulation [6] and chromosome segregation [38]. SPT15 also interacts with several histones (HTB1-2, HTA2), a GTPase involved in actin cytoskeleton organization and vacuolar transport (VPS1) [6], an assembly factor of the INO80 complex (RVB1) [6], and ASF1. ASF1 participates in the assembly of chromatin during DNA replication, the disassembly and re-assembly of chromatin for the activation/repression of gene transcription, and the repair of DNA damage [39].

Relationships with cell cycle regulation—The CDC28-CLB2 complex is a master hub that coordinates a variety of cell cycle-dependent kinases (Fig. 5c), regulating the functions of many cell cycle-linked proteins that have relationships with NER and transcription.

The [CAK1, CDC28, KIN28, CDC37] module represents a kinase cascade that starts with the cyclin-dependent kinase-activating kinase CAK1, leading to the phosphorylation of CDC28 and KIN28 [40]. Defective KIN28 was reported to impair transcription-coupled but not global genome NER [41]. On the other hand, CDC37, a co-chaperone of the heat shock protein HSP82, plays a critical role in regulating cyclin-dependent kinases through stress-activated MAPKKK cascades (summarized in SGD).

Two other cell cycle-regulated kinases, IPL1 and SAK1, participate in two signaling pathways leading to NER-transcription functions, as shown in the left panel of Figure 5c. The pathway with the Aurora kinase IPL1 - a regulator of kinetochore-microtubule attachments [42] - includes RVS167, which uses its SH3 binding domain to mediate the regulation of actin cytoskeleton and cell viability following stress [43]. The other pathway involves SAK1, an activator of the SNF1-SNF4 kinase complex that participates in cellular response to stress and transcription regulation [44]. This pathway also includes the transcription elongation factor ELC1 which inhibits the degradation of SNF4 [45].

Three subunits of TFIIH - SSL1, TFB1, TFB4 - interact with CDC27 (Fig. 4a), an essential component of the anaphase promoting complex (APC). In the identified module [CDC28, CDC5, APC9, CDC16, CDC27], the latter three proteins are core units of APC that cooperates with the Polo-like kinase CDC5 to regulate exit from mitosis [46]. Another TFIIH subunit, SSL2, interacts with the kinase PKC1 activating a MAP kinase cascade - BCK1, MKK1 (highly sensitive to UV [9]), MKK2, and SLT2 - for the regulation of cell growth and cell wall integrity [6].

Discussion

Navigating the landscape—The tremendous power of yeast genetics has enabled determination of the contribution to cell survival of nearly every non-essential gene after challenge with UV [9]. High-throughput and detailed analyses of protein-protein physical interactions have generated large databases (e.g., SGD, BIOGRID) that can be mined to clarify the topology and architecture of important biological processes such as DNA repair. The landscape of protein-protein interactions determined here using computational and statistical methods provides a systematic view of the organization of the NER system that protects against a ubiquitous environmental carcinogen (solar UV radiation causes over one million new cases of skin cancer in the U.S. yearly). Because the biochemistry of DNA repair, replication and transcription is highly conserved from yeast to man, this interactome landscape based on yeast protein-protein physical interaction data also provides a detailed working model of the human NER system.

The network of protein interactions that defines the NER landscape was determined automatically through computational and statistical analyses. We applied set theory instead of graph theory to analyse protein interaction overlaps. Set theory offered some advantages. It permitted the use of recursive techniques that incrementally incorporate protein neighborhoods. Rather than being confined to two neighborhoods imposed by adjacency matrices, recursion enables systematic computation of overlaps for as many neighborhoods as needed until no overlap is found.

The notion of measuring topological overlaps between direct neighbors (one hop) of two genes was defined via adjacency matrices and applied in a graph theory setting [10]. This measure was extended using a set theory interpretation to accommodate more distant neighbours (two or more hops) [47], and direct neighbors of multiple genes (instead of two genes) [48]. While the latter approach was applied to predict genes related to target genes, the former two approaches computed topological overlap measures which were subsequently fed into hierarchical clustering algorithms to find modules. In contrast, we directly computed overlaps among one-hop neighborhoods of multiple proteins, statistically scored the results, and used the significant overlaps to reveal both protein modules and hubs, without requiring clustering algorithms. However, with recursion, computer memory usage increases with the number of neighborhoods. When this number is large (e.g., for the entire yeast interactome), parallel algorithms to partition computation workloads across machines will be needed.

We applied the graph theory algorithm [10] to identify modules among the core list of NER proteins with results comparable to those obtained using the set-theory-based approach. Both methods identify the large TFIIH and RFC modules and the smaller DNA polymerase and chromatin remodelling modules. Both methods of analysis also require manual curation to further group the various modules into a topological landscape with biological meaning.

The core elements of NER as represented in Figure 3 are highly interconnected to accomplish the individual steps of NER, leading to excision of an oligonucleotide containing the DNA photoproduct [5]. The system of protein interactions is consistent with all known steps of damage excision in *S. cerevisiae* and humans with one notable exception. While MSH2 is known to contribute to cell survival after UV in yeast, its contribution in mammalian cells has not been established. The [RAD7, RAD16, ELC1, ABF1] subnetwork that was not included in the initial list of human NER genes appears to couple recognition of DNA damage by the RAD14-RAD4-RAD23 complex to regulation of RNA transcription and protein ubiquitination, thereby spreading the cellular response to DNA damage to other biological processes beyond NER (Fig. 5). RAD16 is homologous to human HLTF (alias HIP116) which has a domain that may bind stalled replication forks [49].

Inclusion of multiple yeast homologs of human NER proteins was useful in revealing a larger subset of related proteins. For example, the human Cockayne syndrome ERCC6 (alias CSB) protein had 16 yeast homologs (Table 1). Although RAD26 was the main homolog, many other homologs (e.g., CHD1, INO80, ISW1-2, MOT1, RAD16, STH1) were components of identified modules and hubs that had helicase and chromatin remodeling functions associated with NER and transcription.

The NER landscape also provides insights into various related pathways and helps formulate new hypotheses. For example, a pathway related to cell cycle regulation (Fig. 5c, left panel) includes IPL1, which was reported to phosphorylate RVS167. This protein binds SYF1, a homolog of the human XPA binding protein XAB2. Together with the interaction between XAB2 and the Cockayne protein CSB [50], these data suggest a signaling cascade initiated from CDC28-CLB2 leading to transcription-coupled NER. RFC2-5 interacts with CTF18 to form a cohesin-loading complex. As UV light induces homologous recombination between cohesed daughter chromatids (sister chromatid exchange) will be of interest to determine whether cohesin loading influences NER. Another example of this type derives from the interaction between the TFIIH module and Cdc27, a member of the anaphase-promoting complex which ubiquitylates proteins to regulate mitosis (Fig. 4a). A key word search using ubiquitin, TFIIH and repair discovered a paper by Nouspikel and Hanawalt [51] describing how levels of the E1 ubiquitin-loading factor may interact with TFIIH to reduce NER in terminally differentiated cells. The value of the computational model was realized, as the presence of protein interactions pointed to a novel biological interaction.

In summary, we demonstrated the application of a set-theory based recursive approach to analyse protein interaction networks and construct a topological and functional landscape for NER. The landscape integrated different pathways manifested through modules and hubs to provide systems-level views of the relationships among NER, transcription, and other biological processes. We took advantage of set theory operations provided by the GNU C++ object-oriented standard template library [15] to build a prototype for both the computational and statistical models. With increasing numbers of protein interactions from large-scale experiments, systematic computationally-driven methods that can automatically identify biologically relevant protein entities will facilitate the analysis of cellular mechanisms in response to DNA damage and other forms of stress.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank Keziban Űnsal-Kaçmaz and Marila Cordeiro-Stone for helpful discussions and ideas and Dan Reed for initial support of this work. Nancy Tran was supported by the Leon and Bertha Goldberg Fellowship. Funding was also provided by PHS grants (ES014635, ES011391, ES010126, CA081343, GM070335, CA074015) and NSF grants to the National Center for Supercomputing Applications (CA SCI-0525308 and CSA SCI-0438712).

References

1. Kolodner RD, Putnam CD, Myung K. Maintenance of genome stability in Saccharomyces cerevisiae. Science 2002;297:552–557. [PubMed: 12142524]

2. Friedberg, EC., et al. DNA Repair and Mutagenesis. Vol. 2nd ed. ASM Press; Washington, D.C.: 2006.

3. Ataian Y, Krebs JE. Five repair pathways in one context: chromatin modification during DNA repair. Biochem. Cell Biol 2006;84:490–504. [PubMed: 16936822]

4. Kraemer KH, Lee MM, Scotto J. DNA repair protects against cutaneous and internal neoplasia: evidence from xeroderma pigmentosum. Carcinogenesis 1984;5:511–514. [PubMed: 6705149]

5. Sancar A, Lindsey-Boltz LA, Unsal-Kacmaz K, Linn S. Molecular mechanisms of mammalian DNA repair and the DNA damage checkpoints. Annu. Rev. Biochem 2004;73:39–85. [PubMed: 15189136]

6. Hong, EL., et al. Saccharomyces Genome Database. 2007. http:://www.yeastgenome.org, ftp://ftp.yeastgenome.org/yeast

7. Wood RD, Mitchell M, Lindahl T. Human DNA repair genes. Mutat. Res 2005;577:275–283. [PubMed: 15922366]

8. Altschul SF, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402. [PubMed: 9254694]

9. Begley TJ, Rosenbach AS, Ideker T, Samson LD. Damage recovery pathways in Saccharomyces cerevisiae revealed by genomic phenotyping and interactome mapping. Mol Cancer Res 2002;1:103–112. [PubMed: 12496357]

10. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi AL. Hierarchical organization of modularity in metabolic networks. Science 2002;297:1551–1555. [PubMed: 12202830]

11. Pawson T, Nash P. Protein-protein interactions define specificity in signal transduction. Genes Dev 2000;14:1027–1047. [PubMed: 10809663]

12. Benjamini Y, Hochberg Y. Controlling the false discovery rate -- a practical and powerful approach to multiple testing. J. Roy. Stat. Soc., Ser. B 1995;57:289–300.

13. McLachlan, GJ.; Do, K-A.; Ambroise, C. Analyzing microarray gene expression data. Wiley-Interscience; Hoboken, N.J.: 2004.

14. Spirin V, Mirny LA. Protein complexes and functional modules in molecular networks. PNAS 2003;100:12123–12128. [PubMed: 14517352]

15. Stepanov A, Lee M. The Standard Template Library. HPLab. Tech. Report 1995;95

16. Shannon P, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 2003;13:2498–2504. [PubMed: 14597658]

17. Chen L, Madura K. Rad23 promotes the targeting of proteolytic substrates to the proteasome. Mol Cell Biol 2002;22:4902–4913. [PubMed: 12052895]

18. Ng JM, et al. A novel regulation mechanism of DNA repair by damage-induced and RAD23-dependent stabilization of xeroderma pigmentosum group C protein. Genes Dev 2003;17:1630–1645. [PubMed: 12815074]

19. Ramsey KL, et al. The NEF4 complex regulates Rad4 levels and utilizes Snf2/Swi2-related ATPase activity for nucleotide excision repair. Mol Cell Biol 2004;24:6362–6378. [PubMed: 15226437]

20. Ribar B, Prakash L, Prakash S. Requirement of ELC1 for RNA polymerase II polyubiquitylation and degradation in response to DNA damage in Saccharomyces cerevisiae. Mol Cell Biol 2006;26:3999–4005. [PubMed: 16705154]

21. Yu S, Owen-Hughes T, Friedberg EC, Waters R, Reed SH. The yeast Rad7/Rad16/Abf1 complex generates superhelical torsion in DNA that is required for nucleotide excision repair. DNA Repair (Amst) 2004;3:277–287. [PubMed: 15177043]

22. Venditti P, Costanzo G, Negri R, Camilloni G. ABFI contributes to the chromatin organization of Saccharomyces cerevisiae ARS1 B-domain. Biochim Biophys Acta 1994;1219:677–689. [PubMed: 7948025]

23. Keogh MC, Cho EJ, Podolny V, Buratowski S. Kin28 is found within TFIIH and a Kin28-Ccl1-Tfb3 trimer complex with differential sensitivities to T-loop phosphorylation. Mol Cell Biol 2002;22:1288–1297. [PubMed: 11839796]

24. Feaver WJ, et al. Dual roles of a multiprotein complex from S. cerevisiae in transcription and DNA repair. Cell 1993;75:1379–1387. [PubMed: 8269516]

25. Lauder S, et al. Dual requirement for the yeast MMS19 gene in DNA repair and RNA polymerase II transcription. Mol Cell Biol 1996;16:6783–6793. [PubMed: 8943333]

26. Svejstrup JQ. Mechanisms of transcription-coupled DNA repair. Nat Rev Mol Cell Biol 2002;3:21–29. [PubMed: 11823795]

27. Kunkel TA, Erie DA. DNA mismatch repair. Annu Rev Biochem 2005;74:681–710. [PubMed: 15952900]

28. Guzder SN, Sommers CH, Prakash L, Prakash S. Complex formation with damage recognition protein Rad14 is essential for Saccharomyces cerevisiae Rad1-Rad10 nuclease to perform its function in nucleotide excision repair in vivo. Mol Cell Biol 2006;26:1135–1141. [PubMed: 16428464]

29. Budd ME, Campbell JL. The pattern of sensitivity of yeast dna2 mutants to DNA damaging agents suggests a role in DSB and postreplication repair pathways. Mutat Res 2000;459:173–186. [PubMed: 10812329]

30. Eissenberg JC, Ayyagari R, Gomes XV, Burgers PM. Mutations in yeast proliferating cell nuclear antigen define distinct sites for interaction with DNA polymerase delta and DNA polymerase epsilon. Mol Cell Biol 1997;17:6367–6378. [PubMed: 9343398]

31. Shivji MK, Podust VN, Hubscher U, Wood RD. Nucleotide excision repair DNA synthesis by DNA polymerase epsilon in the presence of PCNA, RFC, and RPA. Biochemistry 1995;34:5011–5017. [PubMed: 7711023]

32. Pfander B, Moldovan GL, Sacher M, Hoege C, Jentsch S. SUMO-modified PCNA recruits Srs2 to prevent recombination during S phase. Nature 2005;436:428–433. [PubMed: 15931174]

33. Li X, Li J, Harrington J, Lieber MR, Burgers PM. Lagging strand DNA synthesis at the eukaryotic replication fork involves binding and stimulation of FEN-1 by proliferating cell nuclear antigen. J Biol Chem 1995;270:22109–22112. [PubMed: 7673186]

34. Majka J, Burgers PM. The PCNA-RFC families of DNA clamps and clamp loaders. Prog Nucleic Acid Res Mol Biol 2004;78:227–260. [PubMed: 15210332]

35. Bellaoui M, et al. Elg1 forms an alternative RFC complex important for DNA replication and genome integrity. Embo J 2003;22:4304–4313. [PubMed: 12912927]

36. Franco AA, Lam WM, Burgers PM, Kaufman PD. Histone deposition protein Asf1 maintains DNA replisome integrity and interacts with replication factor C. Genes Dev 2005;19:1365–1375. [PubMed: 15901673]

37. Sprouse RO, Brenowitz M, Auble DT. Snf2/Swi2-related ATPase Mot1 drives displacement of TATA-binding protein by gripping DNA. Embo J 2006;25:1492–1504. [PubMed: 16541100]

38. Hsu JM, Huang J, Meluh PB, Laurent BC. The yeast RSC chromatin-remodeling complex is required for kinetochore function in chromosome segregation. Mol Cell Biol 2003;23:3202–3215. [PubMed: 12697820]

39. Tyler JK, et al. The RCAF complex mediates chromatin assembly during DNA replication and repair. Nature 1999;402:555–560. [PubMed: 10591219]

40. Espinoza FH, et al. Cak1 is required for Kin28 phosphorylation and activation in vivo. Mol Cell Biol 1998;18:6365–6373. [PubMed: 9774652]

41. Tijsterman M, Tasseron-de Jong JG, Verhage RA, Brouwer J. Defective Kin28, a subunit of yeast TFIIH, impairs transcription-coupled but not global genome nucleotide excision repair. Mutat Res 1998;409:181–188. [PubMed: 9875293]

42. Cheeseman IM, et al. Phospho-regulation of kinetochore-microtubule attachments by the Aurora kinase Ipl1p. Cell 2002;111:163–172. [PubMed: 12408861]

43. Bauer F, Urdaci M, Aigle M, Crouzet M. Alteration of a yeast SH3 protein leads to conditional viability with defects in cytoskeletal and budding patterns. Mol Cell Biol 1993;13:5070–5084. [PubMed: 8336735]

44. Nath N, McCartney RR, Schmidt MC. Yeast Pak1 kinase associates with and activates Snf1. Mol Cell Biol 2003;23:3909–3917. [PubMed: 12748292]

45. Hyman LE, et al. Binding to Elongin C inhibits degradation of interacting proteins in yeast. J Biol Chem 2002;277:15586–15591. [PubMed: 11864988]

46. Zachariae W, Nasmyth K. Whose end is destruction: cell division and the anaphase-promoting complex. Genes Dev 1999;13:2039–2058. [PubMed: 10465783]

47. Yip AM, Horvath S. The generalized topological overlap matrix for detecting modules in gene networks. Biocomp 2006:451–457.

48. Li A, Horvath S. The multi-node topological overlap measure for gene neighborhoods analysis. Biocomp 2006:445–450.

49. Iyer LM, Babu MM, Aravind L. The HIRAN domain and recruitment of chromatin remodeling and repair activities to damaged DNA. Cell Cycle 2006;5:775–782. [PubMed: 16627993]

50. Nakatsu Y, et al. XAB2, a novel tetratricopeptide repeat protein involved in transcription-coupled DNA repair and transcription. J Biol Chem 2000;275:34931–34937. [PubMed: 10944529]

51. Nouspikel T, Hanawalt PC. Impaired nucleotide excision repair upon macrophage differentiation is corrected by E1 ubiquitin-activating enzyme. Proc. Natl. Acad. Sci 2006;103:16188–93. [PubMed: 17060614]

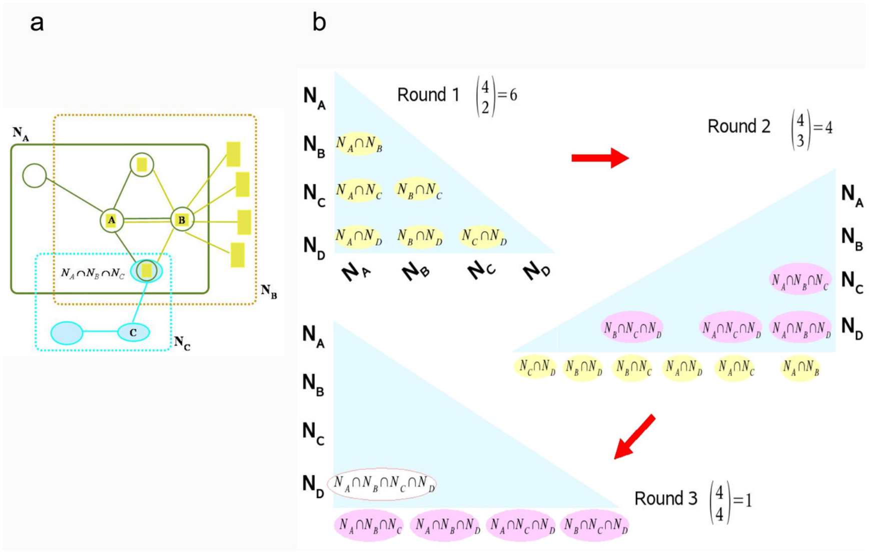

a
b
N_A
N_B
N_C
$N_A \cap N_B \cap N_C$
A
B
C
Round 1 $\binom{4}{2}=6$
N_A N_B N_C N_D
$N_A \cap N_B$
$N_A \cap N_C$ $N_B \cap N_C$
$N_A \cap N_D$ $N_B \cap N_D$ $N_C \cap N_D$
Round 2 $\binom{4}{3}=4$
$N_A \cap N_B \cap N_C$
$N_B \cap N_C \cap N_D$ $N_A \cap N_C \cap N_D$ $N_A \cap N_B \cap N_D$
$N_C \cap N_D$ $N_B \cap N_D$ $N_B \cap N_C$ $N_A \cap N_D$ $N_A \cap N_C$ $N_A \cap N_B$
Round 3 $\binom{4}{4}=1$
$N_A \cap N_B \cap N_C \cap N_D$
$N_A \cap N_B \cap N_C$ $N_A \cap N_B \cap N_D$ $N_A \cap N_C \cap N_D$ $N_B \cap N_C \cap N_D$

Figure 1. An example of systematic recursive computation of protein neighborhood overlaps

(a) *A*, *B*, *C* are the core proteins of 3 neighborhoods N_A,N_B,N_C. Neighborhood N_A is a set of 5 proteins enclosed in the green rectangle. The intersection $N_A \cap N_B \cap N_C$, of the 3 neighborhoods is a set that has a single protein interacting with all 3 core proteins. This intersection's size is at most the size of, the smallest neighborhood. **(b)** Ovals represent neighborhood intersections. Round 1 produces 6 distinct intersections of 2 neighborhoods. Round 2 uses these results to produce 4 distinct intersections of 3 neighborhoods. The final round returns a single intersection of 4 neighborhoods. If any of the intersections is empty, it will be discarded in all subsequent rounds. Empty entries in the matrices represent self-intersections (on the diagonal of the first matrix) or duplicate intersections (due to symmetry and transitivity of set intersections, e.g., $N_A \cap N_B \cap N_C = N_B \cap N_C \cap N_A$).

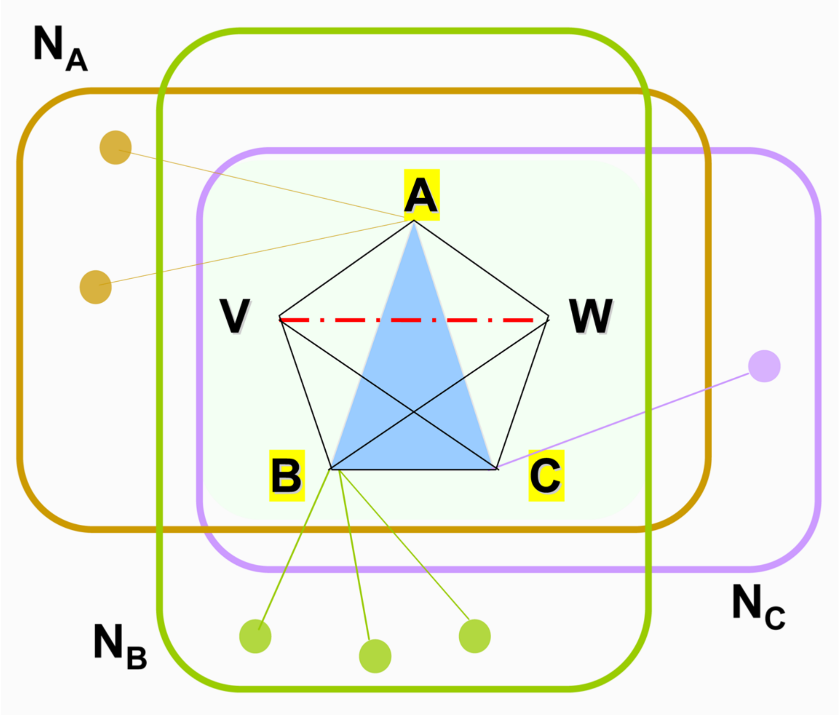


Figure 2. An example illustrating the uncovering of protein modules and aggregates

Core protein set = {*A*, *B*, *C*}, neighborhood intersection set = {*A*, *B*, *C*, *V*, *W*}. Proteins *A*, *B*, and *C* form a module because they are in both sets (rule 1). Applying rules 2 and 3 to incorporate *V* and *W* produces two larger fully-connected modules [*A*, *B*, *C*, *V*] and [*A*, *B*, *C*, *W*]. This new knowledge that *V* and *W* interact with *A, B, and C* predicts a *V*-*W* interaction, resulting in a possibly larger aggregate [*A*, *B*, *C*, *V*, *W*].

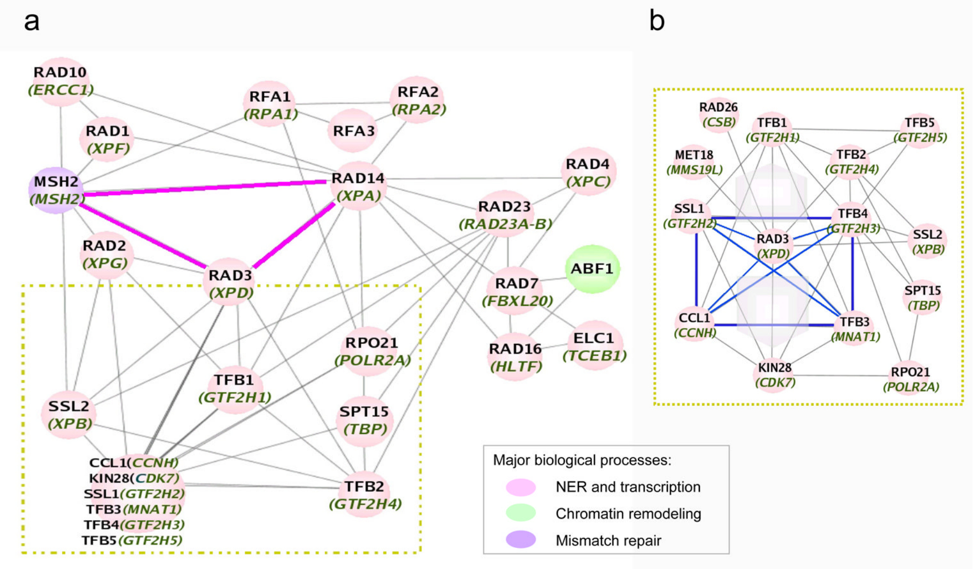
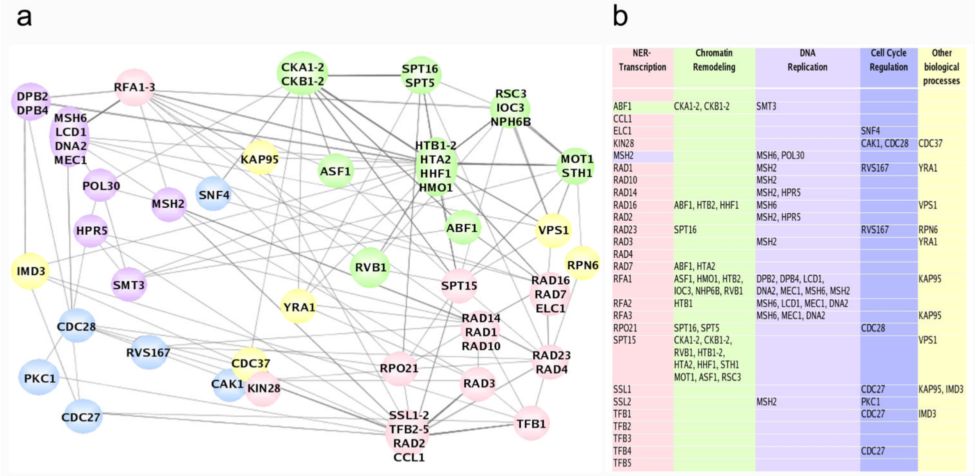


Figure 3. Relationships within the NER-transcription system and organization of TFIIH subunits

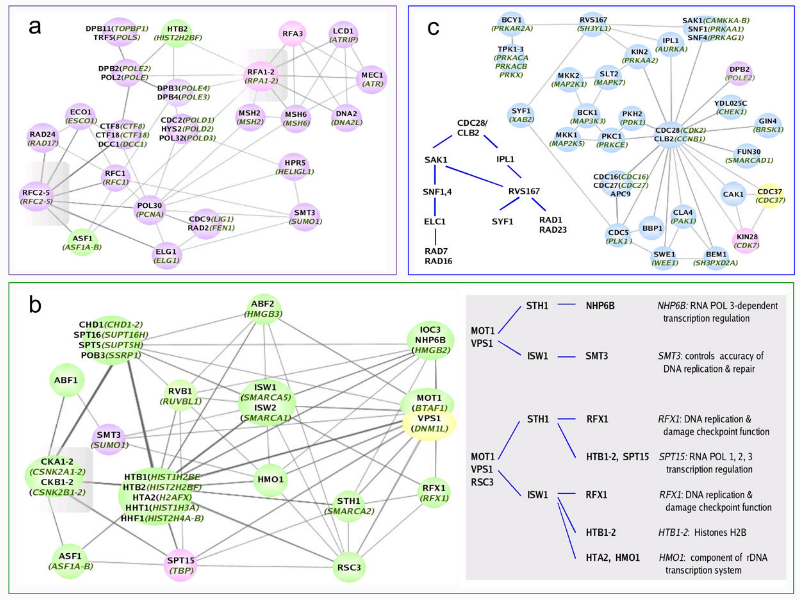
Green italics denote human homologs. **(a)** The NER-transcription system has a hierarchical topology. The top level consists of the [RAD14, RAD3, MSH2] triangle which links to 3 subnetworks: DNA damage sensors that interact with RAD14, TFIIH subunits that interact with RAD3, helicases and nucleases that interact with MSH2. **(b)** Highlighted in blue is a kernel module [RAD3, TFB3, TFB4, CCL1, SSL1] that organizes interactions among TFIIH subunits. Pink-shaded arrows show formation of two larger aggregates emerging from this kernel. Our models did not detect modules or hubs associating with MET18 and RAD26.



NER-Transcription	Chromatin Remodeling	DNA Replication	Cell Cycle Regulation	Other biological processes
ABF1	CKA1-2, CKB1-2	SMT3		
CCL1				
ELC1			SNF4	
KIN28			CAK1, CDC28	CDC37
MSH2		MSH6, POL30		
RAD1		MSH2	RVS167	YRA1
RAD10		MSH2		
RAD14		MSH2, HPR5		
RAD16	ABF1, HTB2, HHF1	MSH6		VPS1
RAD2		MSH2, HPR5		
RAD23	SPT16		RVS167	RPN6
RAD3		MSH2		YRA1
RAD4				
RAD7	ABF1, HTA2			
RFA1	ASF1, HMO1, HTB2, IOC3, NHP6B, RVB1	DPB2, DPB4, LCD1, DNA2, MEC1, MSH6, MSH2		KAP95
RFA2	HTB1	MSH6, LCD1, MEC1, DNA2		
RFA3		MSH6, MEC1, DNA2		KAP95
RPO21	SPT16, SPT5		CDC28	
SPT15	CKA1-2, CKB1-2, RVB1, HTB1-2, HTA2, HHF1, STH1 MOT1, ASF1, RSC3			VPS1
SSL1			CDC27	KAP95, IMD3
SSL2		MSH2	PKC1	
TFB1			CDC27	IMD3
TFB2				
TFB3				
TFB4			CDC27	
TFB5				

Figure 4. A high-level view of relationships among NER-transcription and other biological processes

(a) Nodes (filled circles/ellipses) are color-coded according to biological processes as indicated in (b). Pink nodes represent NER-transcription modules and hubs. Components of the same module are placed within the same node (e.g. RAD14, RAD1, RAD10). All other nodes are direct neighbors of the NER-transcription nodes and represent a hub or a module component (e.g., SNF4 is a component of the [SAK1, SNF1, SNF4] module). **(b)** The first column contains NER-transcription proteins shown in (a). Their direct neighbors participating in other biological processes are shown in other columns.



NHP6B: RNA POL 3-dependent transcription regulation

SMT3: controls accuracy of DNA replication & repair

RFX1: DNA replication & damage checkpoint function

SPT15: RNA POL 1, 2, 3 transcription regulation

RFX1: DNA replication & damage checkpoint function

HTB1-2: Histones H2B

HMO1: component of rDNA transcription system

Figure 5. Detailed views of relationships between NER-transcription and other biological processes
Green italics denote human homologs. Color coding is the same as Fig. 4. Gray-shaded areas highlight the [RFA1-2], [RFC2-5], and [CKA1-2, CKB1-2] kernel modules and their interactions. **(a)** Relationships with DNA replication. **(b)** Relationships with chromatin remodeling. In the right panel, MOT1 and VPS1 constitute a kernel around which a hierarchy of protein interactions are structured. The STH1 and ISW1 ATPases can associate with or without RSC3 to provide a variety of functions along two tree structures. Each tree path from the root to an annotated leaf corresponds to a fully-connected module, e.g., [MOT1, VPS1, RSC3, STH1, RFX1]. **(c)** Relationships with cell cycle regulation. The left panel shows two signaling cascades linking cell cycle regulation with NER.

Table 1

Selection of *Saccharomyces cerevisiae* NER proteins via homologs of human NER genes

This table was constructed from a BLASTP search of yeast homologs of 34 human NER genes [5,7,31]. Protein sequences associated with these genes were compared and BLAST results with similarity p-values <= 10E-012 were selected. Some human NER proteins have several yeast homologs (e.g., CDK7_HUMAN has over 70 homologs), while others have none (RPA1_HUMAN, TF2H5_HUMAN, and DDB1_HUMAN) or weak similarity (DDB2_HUMAN is similar to the yeast splicing factor PRP4 with a p-value of 4E- 007). This table is also supplemented with ten other yeast NER proteins found in SGD. Among these proteins are ABF1, DPB11, POL32, DPB3, RAD7 and ELC1. ABF1, DPB11 and POL32 do not have human homologs. DPB11 appears to be an analog of TOPBP1_HUMAN and POL32 is the third subunit of yeast DNA polymerase delta and functionally similar to POLD3_HUMAN. DPB3 is the third subunit of DNA polymerase epsilon and weakly homologous to POL4_HUMAN, RAD7 is similar to the F-box protein FBXL20_HUMAN, and ELC1 is similar to the transcription elongation factor TCEB1_HUMAN. Mutant phenotypes correspond to deletion mutants used in small scale experiments and high throughput studies. Sensitivity to UV radiation was determined by Begley et al. [9].

	Saccharomyces cerevisiae				
Human NER Protein Name	**Standard Name**	**Systematic Name**	**Mutant Phenotype**	**UV Sensitive 1,2**	**BLAST P-value**
CCNH_HUMAN	CCL1	YPR025C	inviable		1.0E-023
CDK7_HUMAN	ATG1	YGL180W	viable	no	4.0E-016
CDK7_HUMAN	BCK1	YJL095W	viable	no	3.0E-022
CDK7_HUMAN	CAK1	YFL029C	inviable		4.0E-014
CDK7_HUMAN	CDC15	YAR019C	inviable		2.0E-023
CDK7_HUMAN	CDC28	YBR160W	inviable		4.0E-063
CDK7_HUMAN	CDC5	YMR001C	inviable		4.0E-024
CDK7_HUMAN	CHK1	YBR274W	viable	no	9.0E-016
CDK7_HUMAN	CKA1	YIL035C	viable	no	7.0E-029
CDK7_HUMAN	CKA2	YOR061W	viable	no	7.0E-026
CDK7_HUMAN	CLA4	YNL298W	viable	no	4.0E-016
CDK7_HUMAN	CMK1	YFR014C	viable	no	2.0E-015
CDK7_HUMAN	CMK2	YOL016C	viable	no	3.0E-017
CDK7_HUMAN	CTK1	YKL139W	viable	yes	1.0E-052
CDK7_HUMAN	DUN1	YDL101C	viable	yes	2.0E-020
CDK7_HUMAN	FUS3	YBL016W	viable	no	9.0E-045
CDK7_HUMAN	GIN4	YDR507C	viable	no	2.0E-018
CDK7_HUMAN	HOG1	YLR113W	viable	no	9.0E-042
CDK7_HUMAN	HRK1	YOR267C	viable	no	2.0E-014

	Saccharomyces cerevisiae				
Human NER Protein Name	**Standard Name**	**Systematic Name**	**Mutant Phenotype**	**UV Sensitive 1,2**	**BLAST P-value**
CDK7_HUMAN	HSL1	YKL101W	viable	no	8.0E-017
CDK7_HUMAN	IME2	YJL106W	viable	no	6.0E-031
CDK7_HUMAN	IPL1	YPL209C	inviable		3.0E-017
CDK7_HUMAN	KCC4	YCL024W	viable	no	3.0E-019
CDK7_HUMAN	KIC1	YHR102W	viable	no	4.0E-015
CDK7_HUMAN	KIN1	YDR122W	viable	no	9.0E-015
CDK7_HUMAN	KIN2	YLR096W	viable	no	4.0E-016
CDK7_HUMAN	KIN28	YDL108W	inviable		1.0E-077
CDK7_HUMAN	KIN4	YOR233W	viable	no	6.0E-019
CDK7_HUMAN	KIN82	YCR091W	viable	no	7.0E-018
CDK7_HUMAN	KNS1	YLL019C	viable	no	1.0E-016
CDK7_HUMAN	KSS1	YGR040W	viable	no	1.0E-042
CDK7_HUMAN	MCK1	YNL307C	viable	no	1.0E-029
CDK7_HUMAN	MKK1	YOR231W	viable	yes	3.0E-017
CDK7_HUMAN	MKK2	YPL140C	viable	no	1.0E-015
CDK7_HUMAN	MPS1	YDL028C	inviable		2.0E-014
CDK7_HUMAN	MRK1	YDL079C	viable	no	9.0E-039
CDK7_HUMAN	NPR1	YNL183C	viable	no	1.0E-018
CDK7_HUMAN	PBS2	YJL128C	viable	no	9.0E-016
CDK7_HUMAN	PHO85	PHO85	viable	yes	1.0E-062
CDK7_HUMAN	PKC1	YBL105C	inviable		1.0E-017
CDK7_HUMAN	PKH2	YOL100W	viable	no	3.0E-016
CDK7_HUMAN	PKH3	YDR466W	viable	no	4.0E-021
CDK7_HUMAN	PRR2	YDL214C	viable	no	6.0E-015
CDK7_HUMAN	PSK1	YAL017W	viable	no	3.0E-013
CDK7_HUMAN	RAD53	YPL153C	inviable		4.0E-019
CDK7_HUMAN	RIM11	YMR139W	viable	no	1.0E-040
CDK7_HUMAN	RIM15	YFL033C	viable	no	5.0E-016
CDK7_HUMAN	SAK1	SAK1	viable	no	5.0E-016

	Saccharomyces cerevisiae				
Human NER Protein Name	**Standard Name**	**Systematic Name**	**Mutant Phenotype**	**UV Sensitive 1,2**	**BLAST P-value**
CDK7_HUMAN	SCH9	YHR205W	viable	no	2.0E-019
CDK7_HUMAN	SGV1	YPR161C	inviable	no	2.0E-048
CDK7_HUMAN	SKM1	YOL113W	viable	yes	2.0E-015
CDK7_HUMAN	SLT2	YHR030C	viable	no	1.0E-041
CDK7_HUMAN	SMK1	YPR054W	viable	no	8.0E-038
CDK7_HUMAN	SNF1	YDR477W	viable	no	5.0E-025
CDK7_HUMAN	SPS1	YDR523C	viable	yes	5.0E-018
CDK7_HUMAN	SSK2	YNR031C	viable	no	1.0E-025
CDK7_HUMAN	SSK22	YCR073C	viable	no	6.0E-028
CDK7_HUMAN	SSN3	YPL042C	viable	no	6.0E-043
CDK7_HUMAN	STE11	YLR362W	viable	no	1.0E-015
CDK7_HUMAN	STE20	YHL007C	viable	no	9.0E-026
CDK7_HUMAN	TOS3	YGL179C	viable	no	6.0E-020
CDK7_HUMAN	TPK1	YJL164C	viable	no	2.0E-018
CDK7_HUMAN	TPK2	YPL203W	viable	no	3.0E-019
CDK7_HUMAN	TPK3	YKL166C	viable	no	2.0E-019
CDK7_HUMAN	YAK1	YJL141C	viable	no	2.0E-031
CDK7_HUMAN	YBR028C	YBR028C	viable	no	2.0E-018
CDK7_HUMAN	YDL025C	YDL025C	viable	no	2.0E-013
CDK7_HUMAN	YGK3	YOL128C	viable	no	2.0E-023
CDK7_HUMAN	YGR052W	YGR052W	viable	no	2.0E-015
CDK7_HUMAN	YKL161C	YKL161C	viable	no	3.0E-036
CDK7_HUMAN	YPK1	YKL126W	viable	no	2.0E-016
CDK7_HUMAN	YPK2	YMR104C	viable	no	1.0E-016
CDK7_HUMAN	YPL141C	YPL141C	viable	no	1.0E-017
CDK7_HUMAN	YPL150W	YPL150W	viable	no	6.0E-014
CETN2_HUMAN	CDC31	YOR257W	inviable		5.0E-038
CETN2_HUMAN	CMD1	YBR109C	inviable		1.0E-024
CSA_HUMAN	RAD28	YDR030C	viable	no	6.0E-015

	Saccharomyces cerevisiae				
Human NER Protein Name	**Standard Name**	**Systematic Name**	**Mutant Phenotype**	**UV Sensitive 1,2**	**BLAST P-value**
ERCC1_HUMAN	RAD10	YML095C	viable	yes	4.0E-013
ERCC6_HUMAN	CHD1	YER164W	viable	no	7.0E-071
ERCC6_HUMAN	FUN30	YAL019W	viable	no	2.0E-060
ERCC6_HUMAN	INO80	YGL150C	viable	no	3.0E-034
ERCC6_HUMAN	ISW1	YBR245C	viable	no	2.0E-080
ERCC6_HUMAN	ISW2	YOR304W	viable	no	9.0E-082
ERCC6_HUMAN	MOT1	YPL082C	inviable		1.0E-080
ERCC6_HUMAN	RAD16	YBR114W	viable	yes	1.0E-017
ERCC6_HUMAN	RAD26	YJR035W	viable	no	8.0E-166
ERCC6_HUMAN	RAD5	YLR032W	viable	yes	2.0E-014
ERCC6_HUMAN	RAD54	YGL163C	viable	yes	8.0E-070
ERCC6_HUMAN	RDH54	YBR073W	viable	no	2.0E-062
ERCC6_HUMAN	RIS1	YOR191W	viable	no	2.0E-017
ERCC6_HUMAN	SNF2	YOR290C	viable	no	1.0E-074
ERCC6_HUMAN	STH1	YIL126W	inviable		1.0E-077
ERCC6_HUMAN	SWR1	YDR334W	viable	no	8.0E-047
ERCC6_HUMAN	YFR038W	YFR038W	viable	no	8.0E-038
LIG1_HUMAN	CDC9	YDL164C	inviable		9.0D-146
LIG1_HUMAN	DNL4	YOR005C	viable	no	2.0D-026
MAT1_HUMAN	TFB3	YDR460W	inviable		3.0E-031
MMS19l_HUMAN	MET18	YIL128W	viable	yes	2.0E-017
PCNA_HUMAN	POL30	YBR088C	inviable		2.0E-052
RAD23A_HUMAN	RAD23	YEL037C	viable	yes	3.0E-027
RAD23B_HUMAN	RAD23	YEL037C	viable	yes	6.0E-024
RFC1_HUMAN	CTF18	YMR078C	viable	yes	5.0E-020
RFC1_HUMAN	RFC1	YOR217W	inviable		7.0E-114
RFC2_HUMAN	RFC4	YOL094C	inviable		2.0E-108
RFC3_HUMAN	RFC5	YBR087W	inviable		7.0E-079
RFC4_HUMAN	RFC2	YJR068W	inviable		1.0E-090

	Saccharomyces cerevisiae				
Human NER Protein Name	**Standard Name**	**Systematic Name**	**Mutant Phenotype**	**UV Sensitive 1,2**	**BLAST P-value**
RFC5_HUMAN	RFC3	YNL290W	inviable		6.0E-086
RPA1_HUMAN	RFA1	YAR007C	inviable		7.0E-092
RPA2_HUMAN	RFA2	YNL312W	inviable		2.0E-017
TF2H1_HUMAN	TFB1	YDR311W	inviable		9.0E-015
TF2H2_HUMAN	SSL1	YLR005W	inviable		5.0E-071
TF2H3_HUMAN	TFB4	YPR056W	inviable		1.0E-033
TF2H4_HUMAN	TFB2	YPL122C	inviable		2.0E-070
XAB2_HUMAN	SYF1	YDR416W	inviable		6.0E-054
XPA_HUMAN	RAD14	YMR201C	viable	yes	5.0E-014
XPB_HUMAN	SSL2	YIL143C	inviable		0.0E+00
XPC_HUMAN	RAD4	YER162C	viable	yes	2.00E-26
XPD_HUMAN	CHL1	YPL008W	viable	no	1.00E-25
XPD_HUMAN	RAD3	YER171W	inviable		0.00E+00
XPF_HUMAN	RAD1	YPL022W	viable	yes	9.00E-61
XPG_HUMAN	RAD2	YGR258C	viable	yes	3.00E-47
XPG_HUMAN	RAD27	YKL113C	viable	yes	7.00E-14
	ABF1	YKL112W	inviable		
TOPBP1_HUMAN	DPB11	YJL090C	inviable		
POE2_HUMAN	DPB2	YPR175W	inviable		5.0E-038
POE4_HUMAN	DPB3	YBR278W	viable	no	7.0E-006
POLE1_HUMAN	POL2	YNL262W	inviable		0.0E+000
POLD1_HUMAN	CDC2	YDL102W	inviable		0.0E+000
POLD2_HUMAN	HYS2	YJR006W	inviable		2.0E-051
POLD3_HUMAN	POL32	YJR043C	viable	yes	
FBXL20_HUMAN	RAD7	YJR043C	viable	no	1.0E-005
TCEB1_HUMAN	ELC1	YPL046C	viable	no	2.0E-008

Table 2

Summary of results for protein neighborhood intersections

The first two columns are results from the computational model; the last two columns are from the statistical model. Computation stopped at k=16 when all intersections were empty. As the number of neighborhoods increased, the probability of protein sharing decreased (column 3). Among the non-empty intersections shown in column 2, those that were significant and non-redundant were counted in column 4. Note that results for k=14 were redundant and discarded because they were subsets of the intersections found for k=15. Supplementary Table 1 provides a protein list for column 4.

Number of intersecting neighbourhoods (*k*)	Number of non-empty intersections (*t*)	Probability of sharing a protein (p_k)	Number of significant and non-redundant intersections
2	1812	2.8 × 10E-02	29
3	4550	1.3 × 10E-03	41
4	8571	7.6 × 10E-05	31
5	13527	4.8 × 10E-06	21
6	17287	3.0 × 10E-07	28
7	17625	1.8 × 10E-08	16
8	14265	9.6 × 10E-10	9
9	9126	4.6 × 10E-11	5
10	4581	1.9 × 10E-12	5
11	1780	6.8 × 10E-14	3
12	522	2.0 × 10E-15	2
13	110	4.7 × 10E-17	5
14	15	7.7 × 10E-19	0
15	1	6.5 × 10E-21	1