

Genome-Wide Identification and 3D Modeling of Proteins involved in DNA Damage Recognition and Repair (Final Report)

(ID: ER63042-1016196-0006156)

DNA Damage Recognition and Repair (DDR&R) proteins play a critical role in cellular responses to low-dose radiation and are associated with cancer. We have performed a systematic, genome-wide computational analysis of genomic data for human genes involved in the DDR&R process. The significant achievements of this project include:

- 1) Construction of the computational pipeline for searching DDR&R genes, building and validation of 3D models of proteins involved in DDR&R;*
- 2) Functional and structural annotation of the 3D models and generation of comprehensive lists of suggested knock-out mutations;*
- 3) Important Improvement of macromolecular docking technology (see CAPRI results) and its application to predict the DNA-Protein complex conformation;*
- 4) Development of a new algorithm for improved analysis of high-density oligonucleotide arrays for gene expression profiling;*
- 5) Construction and maintenance of the DNA Damage Recognition and Repair Database;*
- 6) Producing 14 research papers (10 published and 4 in preparation).*

1) Construction of the computational pipeline for searching putative DDR&R genes, building and validation of 3D models of proteins involved in DDR&R

An automated computational pipeline to detect new homologues was established using the sequence search protocol. Special scripts automatically access several sequences databases (genomic, EST and protein sequence databases from NCBI, as well as links to informational databanks, like GeneCards, GeneOntology and OMIM) and update our local database servers with new putative DDR&R genes. The maintenance of an in-house mirror is vital to ensure better performance when searching through large amounts of information. Using the weekly updated databases (which include Genbank, Refseq, Ensembl, NCBI-BLAST, STS, Unigene), we perform sequence homologue and regulatory site searches. We analyze these genes before being submitted to the homology modeling pipeline. Newly described DDR&R genes from both our sequence search pipeline and other cases reported in different sequence databases and publications are directed to a second homology model building pipeline. Our methodology makes use of ICM ZEGA-alignment to search for suitable template structures in the RCSB Protein DataBank (PDB). The discrimination factor is given by a probability scoring function that was optimized to properly separate the structurally significant sequence alignments from those that are not structurally correlated.

Homology models were built for all DDR&R proteins that do not have experimentally-solved structures (either by X-ray crystallography or NMR). To do this, an optimized alignment (created by alignSS) was generated when the probability of structural significance indicated a reliable correlation. Models were then built using the ICM homology modeling procedure and refined using different energy minimization and annealing protocols to ensure a realistic placement of all atoms.

Models are not generated for the sake of creating a set of coordinates. They are generated to derive or predict interesting biological information. Therefore, quality-control of the generated models are essential for the next step. Errors can be introduced in experimental phase due to limitations of equipment used (e.g. geometrical errors in covalent geometry, atomic clashes, torsion angles, peptide flip errors and backbone deviations, etc.) and in the model derivation procedure (e.g. tracing the backbones and fitting the side chains, refinement in X-ray crystallography; atom-atom distance determination, structure calculation in NMR models). In the homology modeling procedure, the possible critical errors are: poor template choice, alignment errors, backbone conformational errors, and side chain mis-predictions. Several ICM based scripts have been developed to detect and correct these errors. A normalized structural alignment database (SAD) of 1927 optimal structure-structure alignments has been created to optimize our new alignment method and is open to the public (Figure 1).

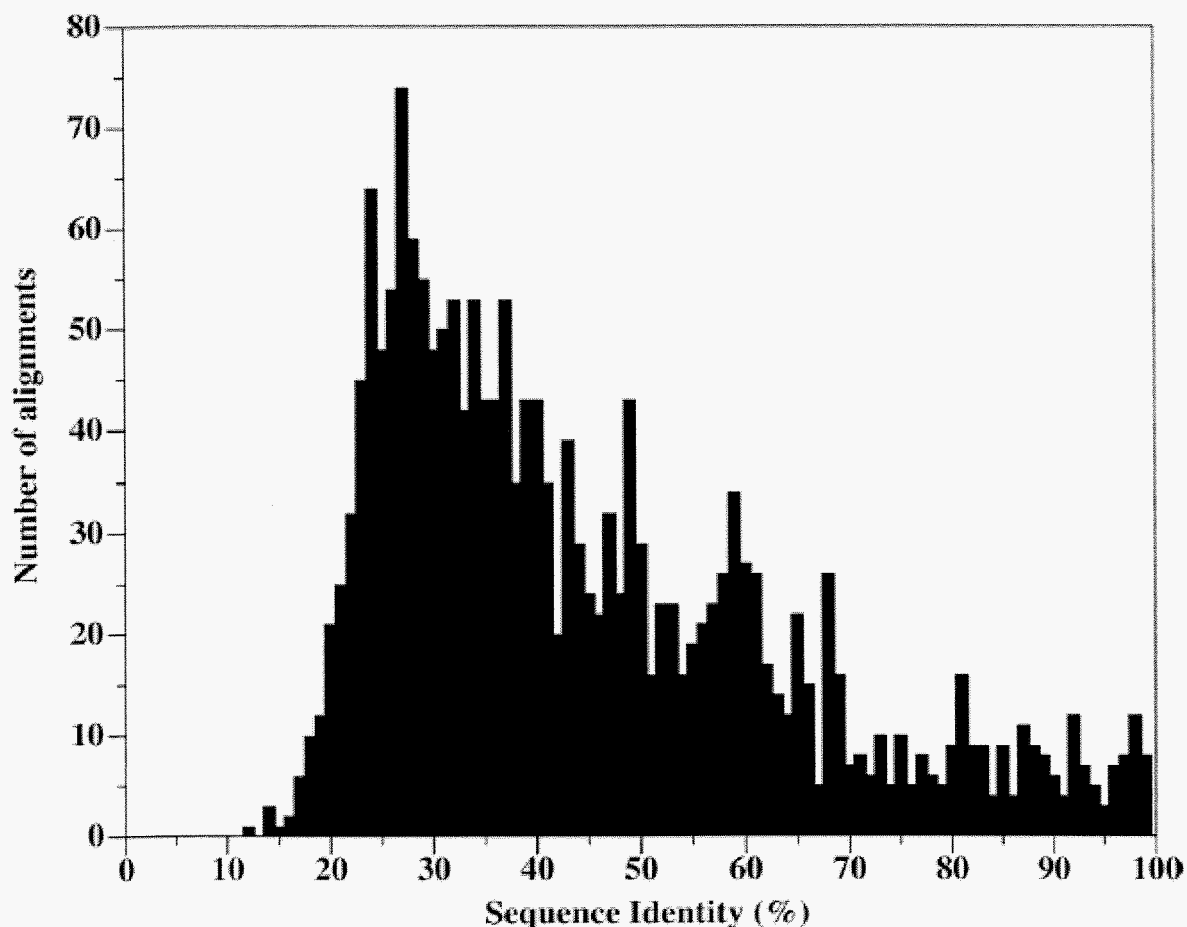


Figure 1: Distribution of sequence identities of alignments in the final SAD database. Despite the majority of SAD structural alignments having sequence identities in the 'twilight-zone' (below approximately 30%), SAD still contains significant amounts of alignments that cover higher ranges of sequence identities.

We have also performed a large-scale loop prediction in internal coordinates using a restrained soft peptide docking algorithm. The test set comprises 805 well-defined loops derived from 410 high-resolution crystal structures. For each loop, a peptide representing the loop was generated, and all free variables were randomized before prediction. The protein surroundings defined as a box stretching 5.0Å in all directions from any loop atom were replaced by grid potentials. Global optimization of energy terms was performed by sampling the conformational space using the biased probability Monte Carlo procedure with local deformation loop movements. The molecular simulation was discontinued upon convergence, defined as the point when three individual simulations reach the same minimum energy (< 1.0 kcal difference). Results were evaluated by comparing the loop conformation to the original PDB structure using static RMSD calculations with superimposed anchor residues. Our results show very accurate prediction for all 805 loops. The average RMSD is 0.66 Å for main chain atoms and 1.35Å for all atoms respectively. The models of DNA repair proteins in our DDR&R database have been rebuilt using the new improved methodology [10,11,12,15].

2) Functional and structural annotation of the 3D models and generation of comprehensive lists of suggested knock-out mutations

When available, the automated pipeline executes the transfer of functional and structural annotations from different homologous proteins to the homology models. However, in several cases, when the DDR&R genes/proteins are totally uncharacterized, alternative strategies are used to detect and suggest residues that may play an important role for the protein's activity. We use three different properties that are usually correlated with an activity and/or function: electrostatics, surface pockets and projection of ligands present in the template structure. For electrostatics, we expect that those DDR&R proteins that directly interact with DNA should present some patches of electropositive potential, helping us to narrow down the range of residues to be considered. For surface pockets, we would expect to find substrates, cofactors and metal ions of enzymes and proteins docked in these pockets. Many times these are critical to the activity of the protein and are also located near functionally important regions. Finally, projection of ligands present in the template structure may help us to find shallow pockets in the homology model that might otherwise be undetectable, as well as adding information to the mapping of cofactor and substrate sites.

To find these three properties in the homology models, specific methodologies and scripts have been developed to enhance the prediction power of this approach. The surface potential electrostatics is calculated using a modified ICM-rebel protocol, designed to filter out smaller spurious patches on the model. A projection script has been written which transfers the ligands from the template to their respective positions on the model, enabling the enhanced prediction of residues that would potentially interact with such ligands. A Large-scale optimization of active site prediction methodology using all known 3D structures has been performed. ICM-Pocket Finder, the program we developed for active site prediction is a combination of geometrical and potential energy consideration. We have optimized the parameters of the method and performed a comprehensive test based on all known structures from the PDB. This method takes a three-

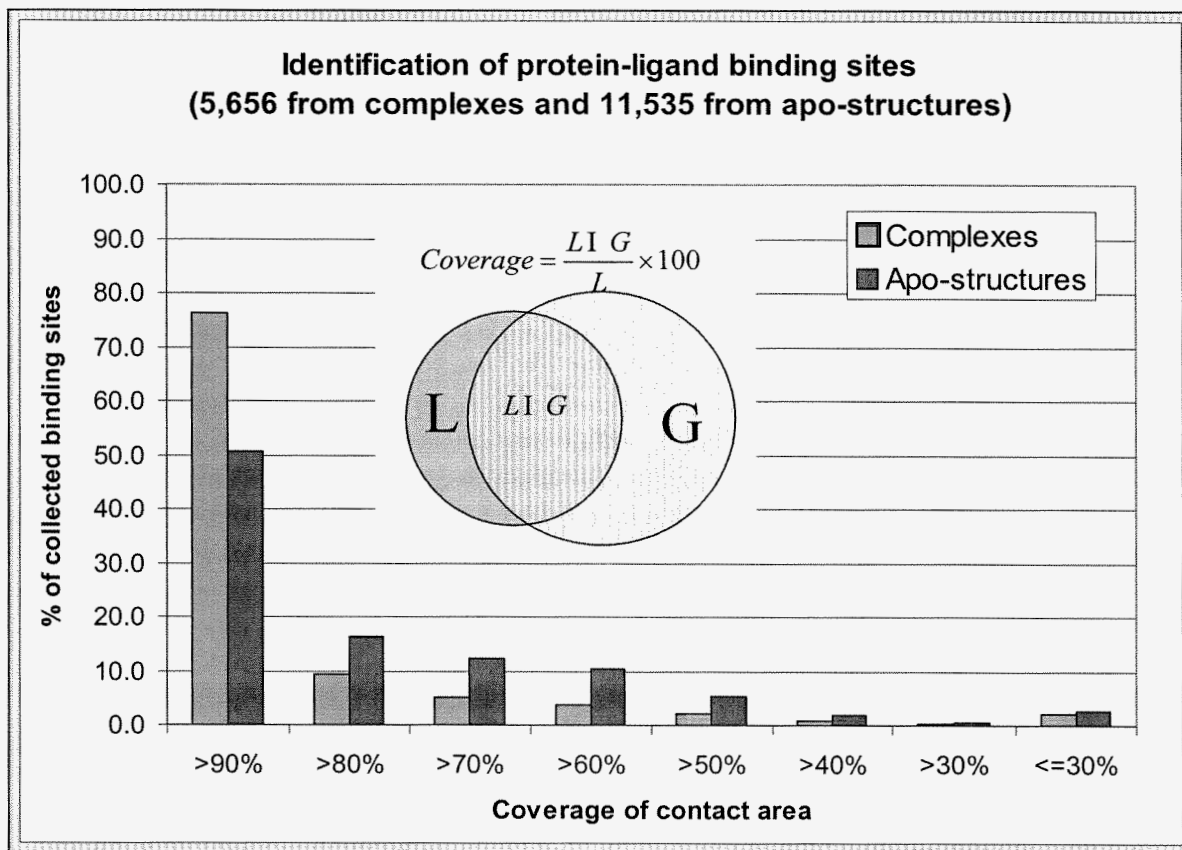


Figure 2. Result of 5,656 binding sites collected from the complex structures and 11,535 binding sites collected from apo-structures.

dimensional protein structure as input and returns the location, volume and shape of the putative binding sites in seconds by using energy potential and without any consideration for a ligand molecule. 17,191 binding sites collected from both complexes and uncomplexed (apo) structures were used to test the method. Of 5,656 binding sites collected from complexes, 98.2% were correctly identified, while the two largest pockets predicted as many as 92.7% of known binding sites. The average ratio of predicted contact area to the total surface area of the protein is 6.8% for the first two pockets. Only in 1.8% of the cases no “pocket density” was found at the ligand location. Further, 11,535 binding sites collected from apo-structures, were predicted with a comparable reliability of 97.8% for all predicted pockets with acceptable volume, and 93.8% for the two largest pockets. The low rate of false negatives and false positives and speed make this method powerful enough to predict protein-ligand binding sites of uncharacterized protein structures (Figure 2).

Partial information from these properties is then combined, yielding a comprehensive table containing different combinations of these properties and their common residues [3,12].

3) Improvement of the protein-protein docking technology and the application to DNA damage repair protein docking

Since DDR&R is a very efficient and flexible system, it is expected that its proteins would form supermolecular assemblies to perform specific tasks during the course of its action. This dynamic

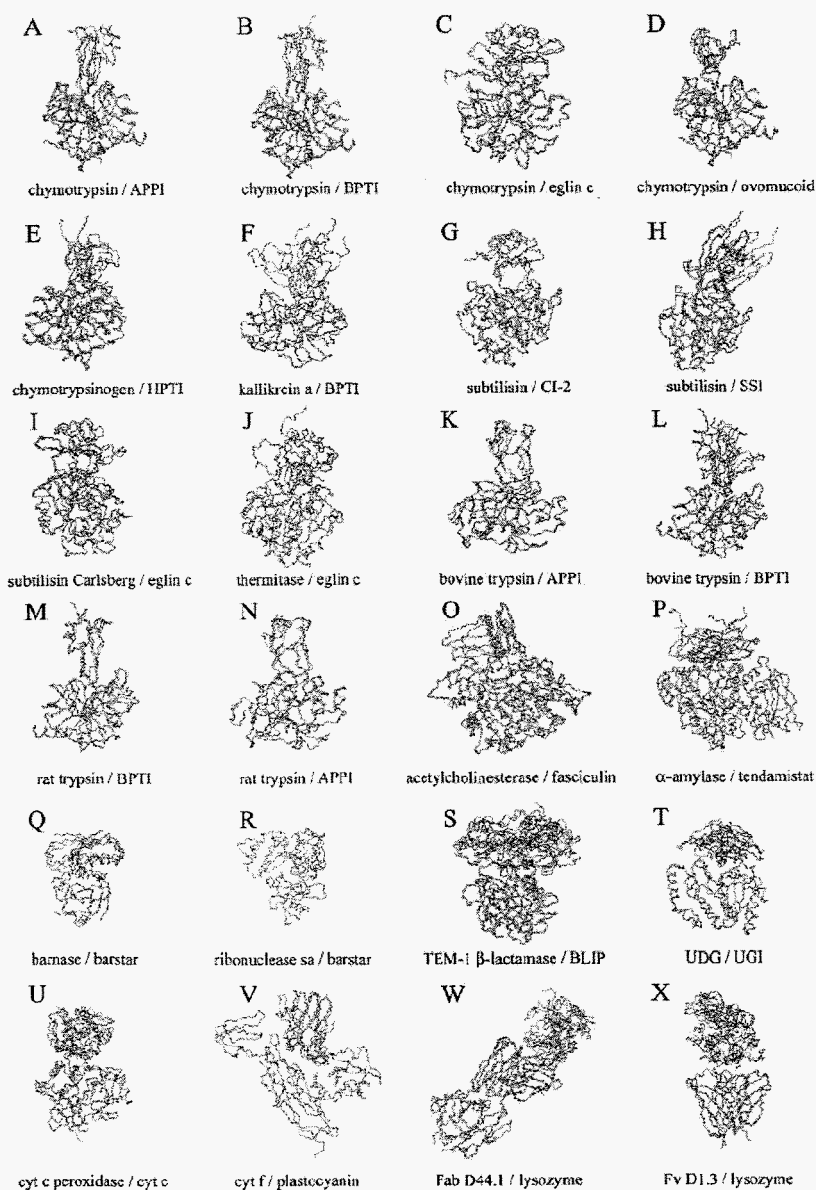


Fig. 3. Representation of the near-native (in red) and lowest energy (in green) solutions after docking unbound subunits and further refinement of the ligand interface side-chains. For those complexes in which the near-native conformation was the lowest energy solution, only the latter is represented (in green). The real ligand structure is represented (in gray) for comparison. Only the receptor C α atoms (in blue) are optimally superimposed onto the corresponding atoms of the real complex (in cyan). For clarity, only backbone atoms (nitrogen, C α , and carbonyl carbon) are shown. The root mean square deviation (RMSD) of the near-native predicted structure respect to the real complex (calculated for the ligand interface C α atoms when the receptor is optimally superimposed onto the real one) is indicated. (A) 1ca0, RMSD 1.2 Å; RANK 1; (B) 1cbw, RMSD 0.7 Å; RANK 1; (C) 1acb, RMSD 4.3 Å; RANK 102; (D) 1cho, RMSD 1.0 Å; RANK 1; (E) 1egi, RMSD 3.1 Å; RANK 12; (F) 2kai, RMSD 5.5 Å; RANK 2; (G) 2sni, RMSD 2.9 Å; RANK 1; (H) 2sic, RMSD 1.9 Å; RANK 7; (I) 1ese, RMSD 2.5 Å; RANK 40; (J) 2tec, RMSD 8.1 Å; RANK 146; (K) 1taw, RMSD 2.9 Å; RANK 1; (L) 2ptc, RMSD 2.0 Å; RANK 3; (M) 3tgi, RMSD 0.8 Å; RANK 1; (N) 1brc, RMSD 1.8 Å; RANK 1; (O) 1fss, RMSD 1.7 Å; RANK 7; (P) 1bvn, RMSD 5.0 Å; RANK 7; (Q) 1bgs, RMSD 4.2 Å; RANK 212; (R) 1ay7, RMSD 6.2 Å; RANK 156; (S) TEM1, RMSD 3.1 Å; RANK 12; (T) 1ugh, RMSD 4.8 Å; RANK 9; (U) 2pcb, RMSD 3.2 Å; RANK 46; (V) 2pcf, RMSD 5.2 Å; RANK 9; (W) 1mlc, RMSD 5.1 Å; RANK 16; and (X) 1vfb, RMSD 3.1 Å; RANK 75. Complexes 1acb, 1ese, 2tec, and 1vfb present ligand backbone deformation on binding (RMSD of the unbound ligand backbone atoms in the interface >1.8 Å with respect to the complexed structure, after optimal superimposition of all ligand C α atoms), so the indicated RMSD values for them may not reflect the accuracy of the predicted near-native conformations.

and precise mechanism, often dictated by the ephemeral and delicate balance of protein-protein interactions play a key role in the recognition and repair processes. In order to better understand these molecular machines, we have developed a protein-protein docking algorithm based on the ICM global energy optimization method. The conformational sampling is achieved with Monte Carlo pseudo-Brownian movements followed by Biased Probability minimization of the interface side-chains. The scoring function is based on soft potentials pre-calculated on a 3-D grid, which helps to improve the speed and efficiency of the minimization procedure.

The docking method has been validated in a set of 24 protein-protein complexes using the unbound subunits (Figure 3), which is the biggest benchmark for protein-protein docking reported so far. The method correctly predicts the near-native conformation within the top 20 solutions in 85% of the cases, and correctly predicts the near-native conformation as the lowest energy solution in 30% of the cases, which clearly outperforms other protein-protein docking methods. [2,3,4,5,6]

Our docking technology was tested in the first blind worldwide competition, the Critical Assessment of PRedicted Interactions (CAPRI), 2003, and received the highest score (*see Table 4, by R. Mendez et al. Proteins 52:51-67*). All-atom refinement of rigid body protein-protein docking solutions using the Biased Probability Monte Carlo Procedure was found to substantially improve the prediction accuracy. A similar refinement procedure that used an all-atom model of the receptor rather than grid-based potentials was also found to improve peptide-protein binding geometry and affinity predictions. Force field parameters such as charges on phosphate groups and the list of torsion angles sampled by the Monte Carlo procedure were optimized such that the refined solutions most accurately reproduce the X-ray structures. Weighting terms for the energy components in a scoring function was fit such that the near-native structure has the lowest score out of all conformations accumulated in the Monte Carlo simulation conformational stack.

Analysis of the docking landscapes generated by docking simulations also permitted the identification of preferred interaction areas on protein surfaces. A new optimized energy function has been validated in a set of 24 protein-protein complexes, and a binding site prediction method has been developed. The predicted sites covered more than 70% of the real interfaces in 71% of the receptors and 54% of the ligands [6].

However, most DNA damage repair proteins interact with both other proteins and DNA molecules. Therefore, as a test case, we applied our efficient protein-protein docking algorithm to a protein-protein interaction that requires the participation of a DNA molecule in the process. We chose a well established system (DNA polymerase Beta – DNA - XRCC1), involved in the repair of DNA single strand break damage. The structures of both proteins were independently solved: the N-terminal domain of the DNA repair protein XRCC1 was deposited in the Protein Databank with the PDB_ID **1xna**, and the DNA molecule bound to DNA polymerase beta (Pol-beta) with PDB_ID **1bpy**. The interacting residues in the complex have been identified using NMR chemical shift mapping (*Marintchev et al 1999, Nature Structural Biology Vol. 6 pp. 884-893 and Gryk et al 2002, Structure Vol. 10, pp. 1709-1720*). We docked these two structures without using any restraints derived from experimental data. The experimental data were used only for cross-validation purposes after docking.

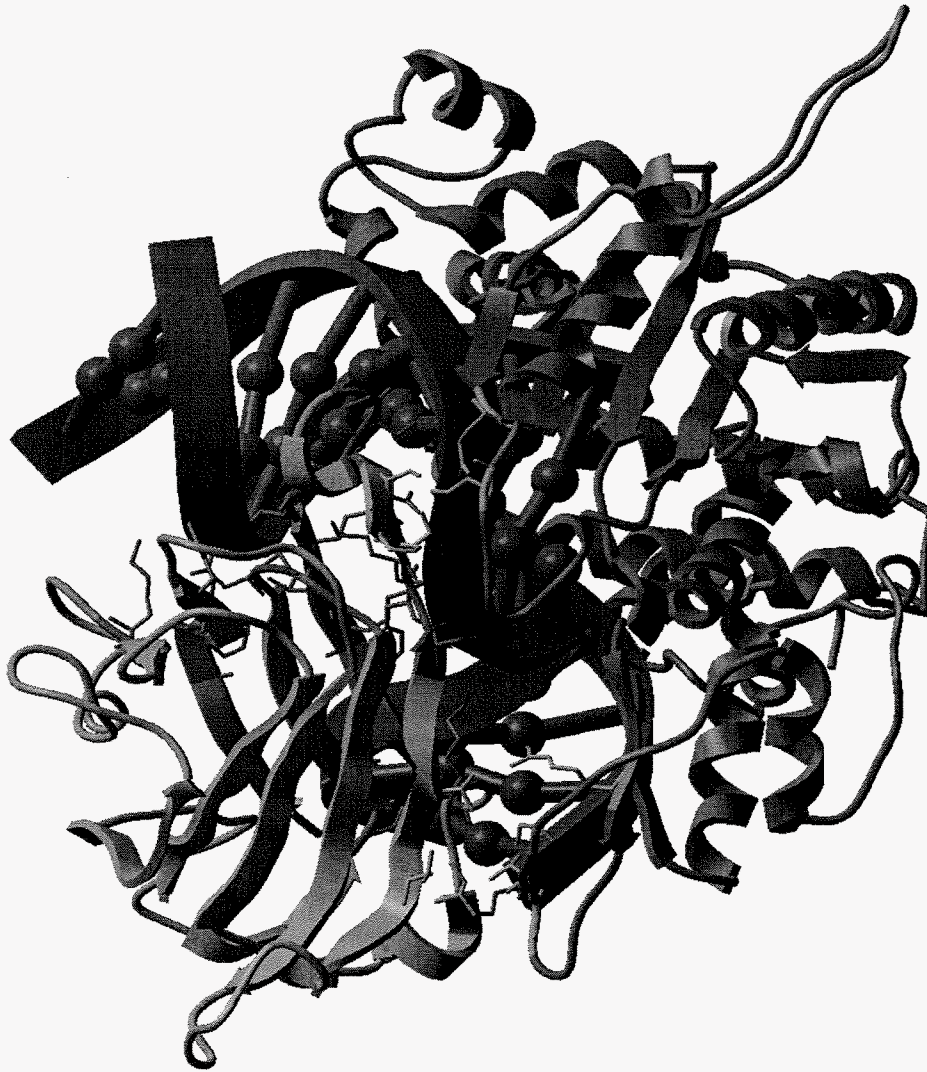


Figure 4. Predicted complex between XRCC1 -NTD (yellow), DNA (red) and Beta-polymerase (blue). The residues predicted to be involved in interactions between XRCC1 and DNA are shown as magenta sticks whilst those that are in contact range intermediating XRCC1 and beta-polymerase interaction are shown as cyan sticks.

Our docking results were inspected visually to identify the correct binding conformation based on the suggested model by previous studies (*Marintchev et al 1999, Nature Structural Biology Vol.6 pp884-893 and Gryk et al 2002, Structure Vol.10, pp1709-1720*). The second docking conformation in the solution stack was the best one which has the largest agreement with the suggested model. The solution has placed the DNA repair protein XRCC1 in a position that it interacts with both the DNA molecule and the polymerase beta. This is the first result of this kind and even though it needs to be further improved, a high rank (#2 out of 280 energy ranked solutions) of the correct solution is an important milestone (Figure 4). This prediction gives new

insight into the interaction of this DDRR complex which is critical for understanding its function [14].

The geometrical complementarity among the elements is one of the main features of this assembly. As reported before, the structure of Pol-beta seems to favour the docking of DNA molecules with single-break damages through the adoption of a specific conformation, creating a well defined cleft on its surface. The docking solution also placed XRCC1 in agreement with the expected shape complementarity, especially that represented by its interaction to the major groove of the DNA molecule through the beta-strand F and the EF loop. Complementarity can also be observed in the distribution of electrostatic patches. In the present case, the major player is the highly electronegative DNA molecule. The complementarity here is extensive and certainly stable, albeit probably not sequence specific. Both proteins have large electropositive patches: XRCC1 has its main patches on the H'X platform and the protruding blade formed by the strand F. Pol-beta presents one very extensive electropositive area that covers the whole DNA-binding face, which is where we can find the cleft. Therefore, the electropositive patch of Pol-beta is further extended by the binding of XRCC1, that interacts with Pol-beta through hydrophobic interactions. The result is that 2/3 of the circumference of electronegative DNA molecule is now surrounded by pol-beta and XRCC1 molecules, that together forms a large, very well defined docking cleft. The results showed interesting and encouraging results for the use of computational methods to predict macromolecular interactions with a DNA molecule.

We have also applied our protein-protein docking methodology to DNA mismatch repair mechanism (MMR). MMR is responsible for the maintenance of DNA fidelity upon replication, and known to be a multicomponent mechanism, including MutS, MutL, MutH, polymerase, ligase, etc. The exact functions and pathway of these proteins are still elusive. Lack of experimental evidence motivated us to predict a partial MMR complex. The crucial proteins in MMR are MutS and MutL, because they, or their homologues, are found in virtually all life forms and they perform the first initiation steps to repair DNA mismatch damages. Thus investigation of interaction between MutS and MutL would be the first step to characterize this complex pathway. So far, the X-ray structures of MutS and MutL have been characterized for *E. Coli* and *Thermus aquaticus* (TAQ): *E. Coli* MutS (1E3M), TAQ MutS (1EWR and 1EWQ), and *E. Coli* MutL (1B62 and 1B63). 1E3M and 1B63 structure were chosen to perform a protein-protein docking simulation to predict the structure of *E. Coli* MutS-MutL complex, find their interaction, and furthermore track down the pathway of DNA mismatch repair. For the X-ray structure of MutL, though a successful crystallization of whole protein has not reported, only N-terminal 40kDa ATPase fragment was identified. Despite of the C terminal 30kDa fragment, the N-terminal 40kDa of MutL (LN40) is able to replace MutL in activating MutH. It indicates that the LN40 contains all of the elements necessary for interacting with DNA and for activation of MutH in the presence of MutS and ATP. But considering the importance of the overall structural alignment of MutL homodimer, there is still doubt as for whether the dimeric form of LN40 crystal represents the MutL dimer reasonably. Although the answer to the question will be difficult to be addressed until the crystal structure of entire MutL dimer is reported, the earlier report suggested that the surface potential of a groove of LN40 dimer is highly positive, making it a prime candidate for DNA binding, and the C-terminal region in MutL, which is absent in the crystal structure, would seal and convert this groove to a hole as in DNA gyrase. The ligand protein, the dimer form of MutL, was generated by applying crystal symmetry using ICM operators to the 1B63 structure.

The complex structure was predicted by side chain refinement after rigid body docking with ICM. The complex structure is the lowest energy conformation, which is 1st rank after refinement. 2nd and 3rd rank conformations also have very similar structure to the 1st ranked one. The conformation with the lowest energy is very similar to the earlier proposed complex structure (Ban et al EMBO J (1998) 17: 1526-1534) [16].

4) Development of a new algorithm for improved analysis of high-density oligonucleotide arrays for gene expression profiling

High-density oligonucleotide arrays have become a valuable tool for high-throughput gene expression profiling. Ultimately we would like to use this technology to identify new DDRR genes. Increasing the array information density and improving the analysis algorithms are two important computational research topics. We have developed a new algorithm, named MOID (Match-Only Integral Distribution), to analyze high-density oligonucleotide arrays. Using known data from both spiking experiments and no-change experiments performed with affymetrix GeneChipO arrays, MOID and the Affymetrix algorithm implemented in Microarray Suite 4.0 (MAS4) were compared. While MOID gave similar performance to MAS4 in the spiking experiments, better performance was observed in the no-change experiments. MOID also provides a set of alternative statistical analysis tools to MAS4. There are two main features that distinguish MOID from MAS4. First, MOID uses continuous P values for the likelihood of gene presence, while MAS4 resorts to discrete absolute calls. Secondly, MOID uses heuristic confidence intervals for both gene expression levels and fold change values, while MAS4 categorizes the significance of gene expression level changes into discrete fold change calls. The results show that by using MOID, Affymetrix GeneChipO arrays may need as little as ten probes per gene without compromising analysis accuracy [7,8,9].

5) Construction and Maintenance of the DNA Damage Recognition and Repair Database

A homepage with important information about the DDR&R proteins and genes has been constructed and made publicly available (<http://abagyan.scripps.edu/DDRR/>). In this database we offer links to protein models for both those with its structure solved and deposited in the Protein DataBank as well as those homology models built through the automated pipeline. Functional sites have been annotated for all models using the binding pocket prediction procedure. The functional site annotation of the DDRR proteins in the DDRR database was analyzed. Residues that were predicted to surround a binding pocket and, at the same time, carry a significant functional annotation were derived and listed. Graphical and tabulated information about the homology model building are also present, as well as a separated table for each protein with suggested mutations as determined by the properties mapping protocol (Figure 5). It summarizes, in a unified and streamlined way, the results we have obtained so far in the DDR&R project [1].

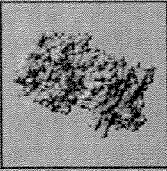
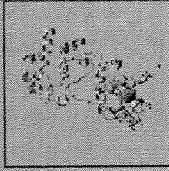
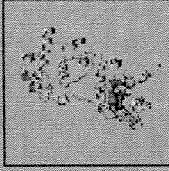

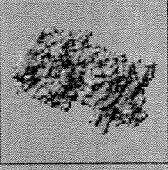
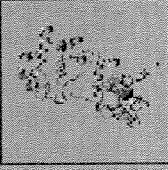

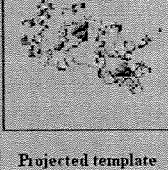
	 Electrostatics	 Putative pockets	 Annotated pockets	 Projected template ligands
 Electrostatics	tyr5, his6, leu7, phe8, arg9, asp10, val11, ala12, glu13, tyr36, ala37, val38, trp39, val40, ser41, glu42, val43, met44, leu45, gln46, gln47, thr48, gln49, val50, ala51, thr52, val53, ile54, asn55, tyr56, tyr57, thr58, gly59, trp60, met61, gln62, lys63, trp64, pro65, thr66, leu67, gln68, asp69, leu70, ala71, ser72, ala73, ser74, leu75, glu76, glu77, val78, asn79, gln80, leu81, trp82, ala83, gly84, leu85, gly85a, tyr85b, tyr85c, ser85d, arg85e, gly92, arg93, arg94, leu95, glu96, glu97, gly98, ala99, arg100, lys101, val102, val103, glu104, glu105, leu106, val150, leu151, cys152, arg153, val154, arg155, ala155a, ile155b, gly155c, ala155d, asp155e, pro155f, ser155g, ser155h, thr155i, leu155j, val155k, ser155l, gln167, gln168, leu169, trp170, gly171, leu172, ala173, gln174, gln175, leu176, val177, asp178, pro179, ala180, arg181, pro182, gly183, asp184, phe185, asn186, gln187, ala188, ala189, met190, glu191, leu192, gly193, ala194, thr195, val196, cys197, thr198, pro199, gln200, arg201, pro202, leu203, cys204, ser205, gln206, cys207, pro208, val209, glu210, ser211, leu212, cys213, arg214	ala12, glu13, trp39, lys101, val102, leu106, val150, leu151, cys152, arg153, val154, arg155, ala155a, ala155d, leu176, phe185, ala189, leu192, val196, cys197, pro202, cys204, cys207, cys204, cys207, val209, cys213, arg214	cys197, cys204, cys207, cys213	thr66, gln68, tyr85c, arg153, val154, val196, cys197, pro202, cys204, cys207, glu210, cys213
 Putative pockets	ala12, glu13, trp39, lys101, val102, leu106, val150, leu151, cys152, arg153, val154, arg155, ala155a, ala155d, leu176, phe185, ala189, leu192, val196, cys197, pro202, cys204, cys207, val209, cys213, arg214	ala12, glu13, thr15, ala16, trp39, lys101, val102, leu106, met110, pro111, thr116, leu117, leu119b, val123, val147, val150, leu151, cys152, arg153, val154, arg155, ala155a, ala155d, leu176, phe185, ala189, leu192, val196, cys197, pro202, cys204, cys207, val209, cys213, arg214, ala215, arg218	cys197, cys204, cys207, cys213	arg153, val154, val196, cys197, pro202, cys204, cys207, cys213, ala215
 Annotated pockets	cys197, cys204, cys207, cys213	cys197, cys204, cys207, cys213	cys197, cys204, cys207, cys213	cys197, cys204, cys207, cys213
 Projected template ligands	thr66, gln68, tyr85c, arg153, val154, val196, cys197, pro202, cys204, cys207, glu210, cys213	arg153, val154, val196, cys197, pro202, cys204, cys207, cys213, ala215	cys197, cys204, cys207, cys213	ala29d, thr66, gln68, tyr85c, his109, gly124, arg125, arg153, val154, val196, cys197, pro202, cys204, cys207, glu210, cys213, ala215, arg216
	Occurs in Electrostatics AND ANY other field	Occurs in Putative Pockets AND ANY other field	Occurs in Annotated Pockets AND ANY other field	Occurs in Projected Ligands AND ANY other field
	ala12, glu13, trp39, thr66, gln68, tyr85c, lys101, val102, leu106, val150, leu151, cys152, arg153, val154, arg155, ala155a, ala155d, leu176, phe185, ala189, leu192, val196, cys197, pro202, cys204, cys207, val209, glu210, cys213, arg214	ala12, glu13, trp39, lys101, val102, leu106, val150, leu151, cys152, arg153, val154, arg155, ala155a, ala155d, leu176, phe185, ala189, leu192, val196, cys197, pro202, cys204, cys207, val209, cys213, arg214, ala215	cys197, cys204, cys207, cys213	thr66, gln68, tyr85c, arg153, val154, val196, cys197, pro202, cys204, cys207, glu210, cys213, ala215

Figure 5: A screen shot of the DDR&R database. The intersections in the table represent residues common to the considered four properties (Electrostatic potential distribution on the surface, putative pockets, annotated binding sites, and projected template ligands).

References:

1. DNA Damage Recognition and Repair database: <http://abagyan.scripps.edu/DDRR/>
2. Fernandez-Recio, J., Totrov, M., and Abagyan, R. (2003). ICM-DISCO Docking by Global Energy Optimization With Fully Flexible Side-Chains. *Proteins* 52:113-117
3. Fernandez-Recio, J., Totrov, M., and Abagyan, R. (2004). Identification of Protein Protein Interaction Sites from Docking Energy Landscapes. *JMB* 335 (3): 843:865
4. Katrich, S., Totrov, M., and Abagyan, R. (2003). ICFF: A new method to incorporate implicit flexibility into an internal coordinate force field. *J. Comput. Chem.* 24:254-265
5. Fernandez-Recio, J., Totrov, M., and Abagyan, R. (2002). Screened charge electrostatic model in protein-protein docking simulations. *Pac Symp Biocomput.* 2002. 552-563
6. Fernandez-Recio, J., Totrov, M., and Abagyan, R. (2002). Soft protein-protein docking in internal coordinates. *Proteins Science* 11:280:291
7. Zhou, Y. and Abagyan, R. (2002). Match-Only Integral Distribution (MOID) Algorithm for High-Density Oligonucleotide Array Analysis. *BMC Bioinformatics* 3:3
8. Volkman, S.K., Hartl, D.L., Wirth, D.F., Nielsen, K.M., Choi, M., Le Roch, K.G., Abagyan, R., Winzeler, E.A. (2002). Excess Polymorphisms in Genes for Membrane Proteins in *Plasmodium Falciparum*. *Science* 298, 216-218
9. Zhou, Y. and Abagyan, R. (2003). Algorithms for High-Density Oligonucleotide Array. *Current Opinion in Drug Discovery & Development* 6(3):339-345
10. Marsden BD, Abagyan, RA (2003). Identifying errors in three dimensional models, *Protein Structure*, Editor: Daniel Chasman, Marcel Decker Inc, 2003
11. Marsden BD, Abagyan, RA (2004). SAD--a normalized structural alignment database: improving sequence-structure alignments, *Bioinformatics* 2004; DOI: 10.1093/bioinformatics/bth244
12. Structural alignment database: <http://abagyan.scripps.edu/lab/web/sad/show.cgi>
13. Comprehensive identification and prediction of protein-ligand binding sites, Jianghong An, Maxim Totrov and Ruben Abagyan. In preparation.
14. Successful protein-protein docking in the presence of a DNA ligand in the interface, Jianghong An, Wen Hwa Lee and Ruben Abagyan. In preparation.
15. Loop modeling in Internal Coordinates against pre-calculated grid potentials, Therese Enneqvist and Ruben Abagyan. In preparation.

16. The Prediction of E. Coli MutS-MutL Complex Structure by Protein-Protein Docking Simulation, Hyuk Soon Choi and Ruben Abagyan. In preparation.