BMC Bioinformatics

Research article

Interaction profile-based protein classification of death domain

Drew Lett[1], Michael Hsing[2] and Frederic Pio*[2]

Address: [1]Department of Computer Science, Simon Fraser University, 8888 University Drive, Burnaby, B.C. Canada, V5A 1S6 and [2]Department of Molecular Biology and Biochemistry, Simon Fraser University, 8888 University Drive, Burnaby, B.C. Canada, V5A 1S6

Email: Drew Lett - dclett@sfu.ca; Michael Hsing - mmhsing@sfu.ca; Frederic Pio* - fpio@sfu.ca

* Corresponding author

Published: 09 June 2004

BMC Bioinformatics 2004, **5**:75 doi:10.1186/1471-2105-5-75

Received: 07 February 2004

Accepted: 09 June 2004

This article is available from: http://www.biomedcentral.com/1471-2105/5/75

Abstract

Background: The increasing number of protein sequences and 3D structure obtained from genomic initiatives is leading many of us to focus on proteomics, and to dedicate our experimental and computational efforts on the creation and analysis of information derived from 3D structure. In particular, the high-throughput generation of protein-protein interaction data from a few organisms makes such an approach very important towards understanding the molecular recognition that make-up the entire protein-protein interaction network. Since the generation of sequences, and experimental protein-protein interactions increases faster than the 3D structure determination of protein complexes, there is tremendous interest in developing *in silico* methods that generate such structure for prediction and classification purposes. In this study we focused on classifying protein family members based on their protein-protein interaction distinctiveness. Structure-based classification of protein-protein interfaces has been described initially by Ponstingl *et al.* [1] and more recently by Valdar *et al.* [2] and Mintseris *et al.* [3], from complex structures that have been solved experimentally. However, little has been done on protein classification based on the prediction of protein-protein complexes obtained from homology modeling and docking simulation.

Results: We have developed an *in silico* classification system entitled HODOCO (Homology modeling, Docking and Classification Oracle), in which protein Residue Potential Interaction Profiles (RPIPS) are used to summarize protein-protein interaction characteristics. This system applied to a dataset of 64 proteins of the death domain superfamily was used to classify each member into its proper subfamily. Two classification methods were attempted, heuristic and support vector machine learning. Both methods were tested with a 5-fold cross-validation. The heuristic approach yielded a 61% average accuracy, while the machine learning approach yielded an 89% average accuracy.

Conclusion: We have confirmed the reliability and potential value of classifying proteins *via* their predicted interactions. Our results are in the same range of accuracy as other studies that classify protein-protein interactions from 3D complex structure obtained experimentally. While our classification scheme does not take directly into account sequence information our results are in agreement with functional and sequence based classification of death domain family members.

Background

The genomic revolution has provided vast protein data resources now waiting to be transformed into usable knowledge that can be applied to solve pressing biological problems. Classification remains a favorite method for performing such transformations because of its intuitiveness and robustness against errors. Several schemes have now been proposed for automatic classification of proteins [4,5]. They range from simple amino acid sequence comparisons, through more localized motif-based methods [6-8] further improved by position specific scoring matrices [9] and finally to hidden Markov model profile-based methods [10]. Alternatively, structure-based classification provides a more direct means of inferring function, albeit on the much smaller structural databases [11,12]. Recently, groups have taken an integrated approach that blends the advantages of the methods discussed above [13,14].

The high-throughput generation of protein-protein interaction data from a few organisms has been carried out [15-18]. This wealth of experimental data requires new computational mining approaches to help us understand molecular recognition in protein-protein interaction networks. Since the generation of sequences and experimental protein-protein interactions increases faster than the 3D structure determination of protein complexes, there is tremendous interest in developing *in silico* methods that could predict macromolecular structures and assembly for prediction and classification purpose.

For example, computational approaches based on sequence, expression and literature abstract data have been developed to predict protein-protein interactions [19]. These methods are based on the assumption that non-homologous pairs of genes that show correlated behavior across data from different sources should interact with each other. In addition, structure-based classification of protein-protein interfaces has been described initially by Ponstingl *et al*. [1] and more recently by Valdar *et al*. [2] and Mintseris *et al*. [3], from complex structures that have been solved experimentally.

The last decade has seen enormous progress in the reliability and accuracy of 3D structure-based *in silico* techniques including 3D structure prediction based on sequence homology and macromolecular docking. Competitions in both domains have spurred the ingenuity necessary for tackling these challenging problems [20,21]. In this study we combine these two approaches to perform protein classification.

To efficiently dock two molecules that participate in a protein-protein or protein-ligand interaction, a certain number of steps have to be determined [22]. The process first involves an efficient search and matching algorithm that covers the conformational space, and then one or more selective scoring functions that can eliminate efficiently between native and non-native solutions. Docking algorithms are defined and classified by the extent of flexibility that they attempt to address (1) Rigid body docking, where the two molecules are rigid solid bodies, (2) Semi-flexible docking where one molecule, the receptor, is considered a rigid body while the ligand, generally smaller, is considered flexible, and finally (3) flexible docking where both molecules are considered flexible. Flexible docking is now becoming more popular because it takes into account conformational changes that generally occur when proteins interacts with each other. However, rigid body docking simulation has already been widely employed and used successfully in the docking of protein-protein complexes [22,23]. In this method flexibility can be incorporated through a "soft belt" into which atoms from the second molecule can penetrate, reducing drastically the complexity [22] and increasing the speed of the simulation. Rigid body docking is based on the observation that 3D protein complexes reveal a close geometric match at the interface of a receptor and a ligand. Since many false positives with better scores than the true solution are very often obtained, additional rescoring functions have been introduced to eliminate these wrong solutions [24].

In this study, rigid body docking is applied to the classification of protein-protein interactions in the death domain superfamily. We chose rigid body docking because of its higher speed. The scheme uses *in silico* protein-protein interaction predictions, applied to 3D protein structures built using homology modeling, as its exclusive means of performing classifications. We implemented the approach in a system called HODOCO (HOmology modeling, DOcking, and Classifying Oracle), and used Residue Pair Interaction Profiles (RPIPs) as a means to summarize protein interaction characteristics. The system was applied successfully to the problem of classifying members of the human death domain superfamily. We show that despite the limited reliability of current docking algorithm, interaction profile-based classification of this family can be obtained with 90% accuracy.

Results and Discussion

The goal of this study was to perform protein classification from *in silico* predicted protein-protein interaction. We chose to concentrate on the death domain superfamily as our model family. The superfamily consists of four families: Caspase-associated recruitment domain (CARD), death effector domain (DED), death domain (DD), and pyrin/AIM/ASC/DD-like domain (PAAD). Each domain in the superfamily has a characteristic 6-helix bundle fold

Table 1: Human death domain superfamily members with known structures. PDB codes are followed by chain identifiers; an underscore represents the only chain http://www.rcsb.org.

Family	Gene Name	PDB Code
CARD	APAF1	1c15:A 2ygs:A 3ygs:C 1cy5:A 1cww:A
CARD	CASP9	3ygs:P
CARD	ICEBERG	1dgn:A
CARD	CRADD	3crd:_
DED	FADD	1a1w:_ 1a1z:_
DEATH	FADD	1e3y:A 1e41:A 1fad:A
DEATH	NGFR	1ngr:_
DEATH	TNFRSF6	1ddf:_
DEATH	TNFRSF1A	1ich:A
Total:	**9**	

Table 2: Gene names and RefSeq IDs of the sequences used in this study. Data from the UCSC Genome Browser [42] April 2003 assembly.

CARD		DED		DD		PAAD	
Gene Name	**RefSeq ID**	**Gene Name***	**RefSeq ID**	**Gene Name**	**RefSeq ID**	**Gene Name**	**RefSeq ID**
APAF1	NM_001160	CASP10(19)	NM_001230	ANK1	NM_000037	AIM2	NM_004833
ASC	NM_013258	CASP10(114)	NM_001230	ANK2	NM_001148	ASC	NM_013258
BCL10	NM_003921	CASP8(2)	NM_001228	ANK3	NM_001149	CIAS1	NM_004895
BIRC2	NM_001166	CASP8(100)	NM_001228	CRADD	NM_003805	DEFCAP	NM_014922
BIRC3	NM_001165	CFLAR(1)	NM_003879	DAPK1	NM_004938	MEFV	NM_000243
CARD10	NM_014550	CFLAR(92)	NM_003879	FADD	NM_003824	NALP10	NM_176821
CARD11	NM_032415	DEDD	NM_004216	IRAK1	NM_001569	NALP2	NM_017852
CARD12	NM_021209	DEDD2	NM_133328	IRAK2	NM_001570	NALP8	NM_176811
CARD14	NM_024110	FADD	NM_003824	IRAK3	NM_007199	PYC1	NM_152901
CARD6	NM_032587	PEA15	NM_003768	LRDD	NM_018494	PYPAF4	NM_134444
CARD9	NM_022352			MYD88	NM_002468	PYPAF5	NM_138329
CASP1	NM_001223			NGFR	NM_002507	RNO2	NM_033297
CASP2	NM_001224			RIPK1	NM_003804		
CASP4	NM_001225			THOC1	NM_005131		
CASP5	NM_004347			TNFRSF10A	NM_003844		
CASP9	NM_001229			TNFRSF10B	NM_003842		
CRADD	NM_003805			TNFRSF1A	NM_001065		
DEFCAP	NM_014922			TNFRSF21	NM_014452		
ICEBERG	NM_021571			TNFRSF25	NM_003790		
RIPK2	NM_003821			TNFRSF6	NM_000043		
TUCAN	NM_014959			TRADD	NM_003789		
Total:	**21**		**10**		**21**		**12**

*Gene names followed by parentheses have more than one death effectors domain. The number in parentheses is the first amino acid of the domain according to Pfam release 10.0.

called the "death fold", and performs protein-protein interactions between members of the same subfamily class. Interaction between superfamily members has been shown to be a functional mechanism of signal transduction in apoptosis (programmed cell death), and therefore is key to many vital life processes such as the proper maintenance of homeostasis and the immune system, as well as diseases such as neurodegenerative disorders and cancer. The characteristics of the superfamily, such as the low intrafamily sequence identity (30% on average), the high interfamily structural similarity, and the fact that family membership is primarily defined by the ability of members of the same family to interact exclusively with each other [35,36] make it ideal for classification based on interaction profiles. The combined literature survey from the database of Protein FAMily http://pfam.wustl.edu, and an iterative BLAST search resulted in 80 sequences, distributed moderately evenly across families. Nine had

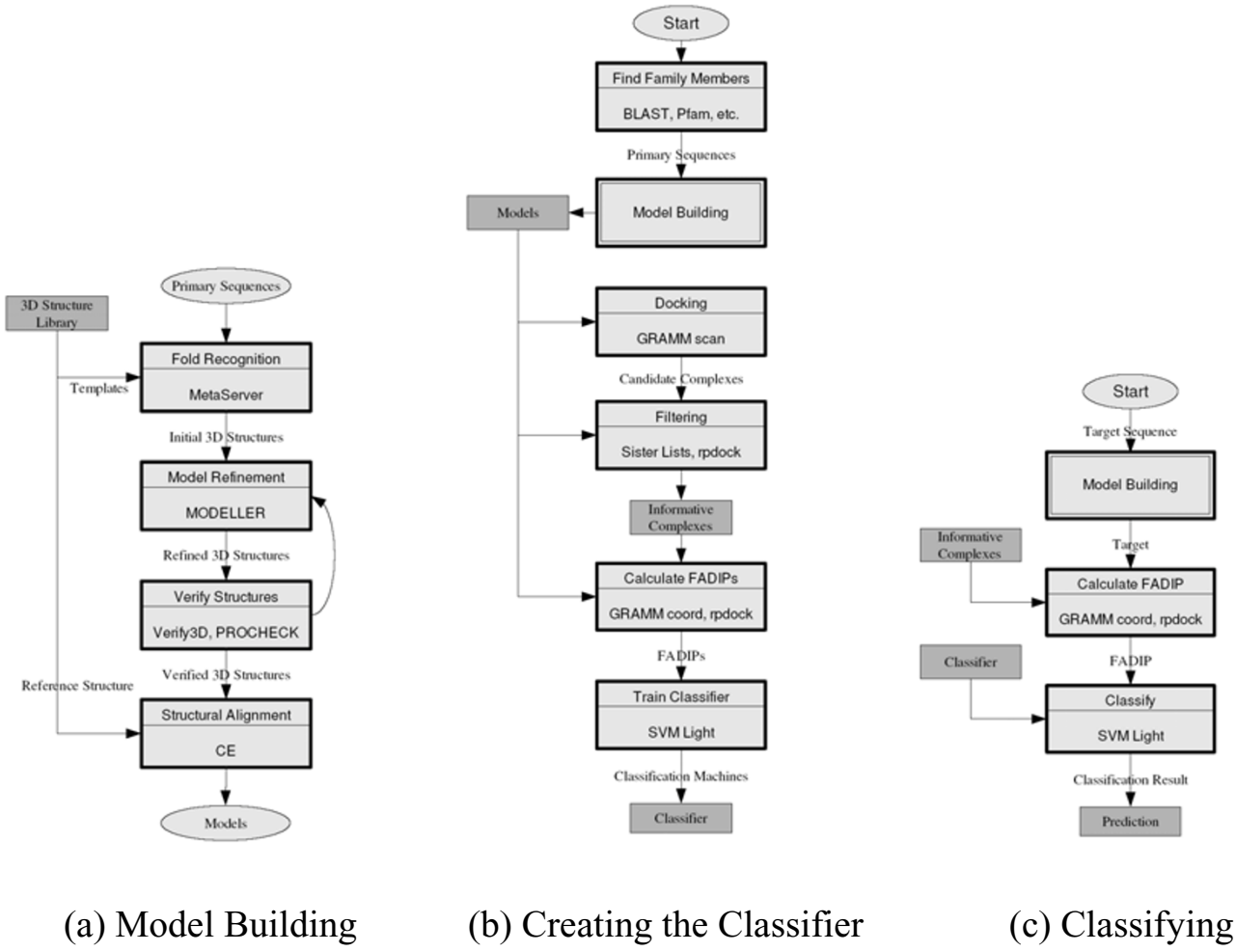

Figure 1

The HODOCO system architecture. (a) The process for determining the missing 3D structures of family members from their amino-acids sequence; (b) The classification engine creation pipeline; and (c) The process for classifying new family members. Light boxes represent computation steps. Within each computation step is a description (above the line) and a list of programs that the step uses (below the line). Dark boxes represent data sources and/or sinks. Edge labels represent the data passed from one step to the next.

solved 3D structures (Table 1). The rest had to be modeled *in silico*, and those that could not be modeled to a reasonable degree of confidence were removed from the study. What remained were 64 models (i.e. $m = 64$): 21 CARD, 10 DED, 21 DD, and 12 PAAD (Table 2).

Great care was taken during the model building process to ensure that the highest quality models were accepted. This degree of caution was required to avoid the risk of propagating inaccuracies throughout the system. GRAMM was chosen as the docking engine for its ability to obtain raw putative complexes that have not been further filtered, thereby allowing us to estimate a signal-to-noise ratio from the raw data and in turn allowing us to devise a procedure for reducing the search space. Docking algorithm was used to build a database that could be mined for specific complexes with properties unique to a given family. We chose to perform docking only between members of the same family as the intrafamily complexes provided a broad enough sampling to yield high accuracy rates and limit our computational cost. Keeping in mind GRAMM's asymmetric algorithm, $21^2 + 10^2 + 21^2 + 12^2 = 1126$ docking simulations were conducted each offering 1000 putative complexes, for a total of over 1.1 million putative complexes.

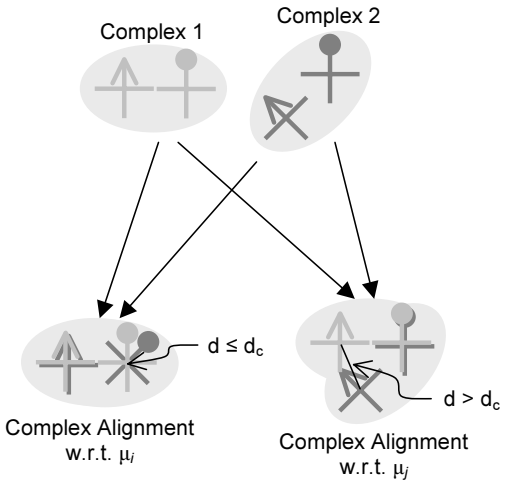

Figure 2
Intersection of sister lists. Complex 1 resulted from docking μ_i against μ_j, $i, j \in [1...m]$. Complex 2 resulted from docking μ_j against μ_i. μ_i is represented as a cross with an arrowhead; μ_j as a cross with a bulb. The different shades of gray are used to distinguish instances of the same model. In this case, if we align the two complexes by superimposing both instances of μ_i, then the distance between the centers of the two instances of μ_j is less than or equal to the cutoff, d_c, so both complexes are retained. Note that the instances of μ_j are not superimposed; only their centers coincide. Here, μ_i is the aligned model and μ_j is the orientation-independent model. When we align the two complexes by superimposing both instances of μ_j, then the distance between the centers of μ_i is greater than the cutoff. However, once a complex is retained, it is never again rejected.

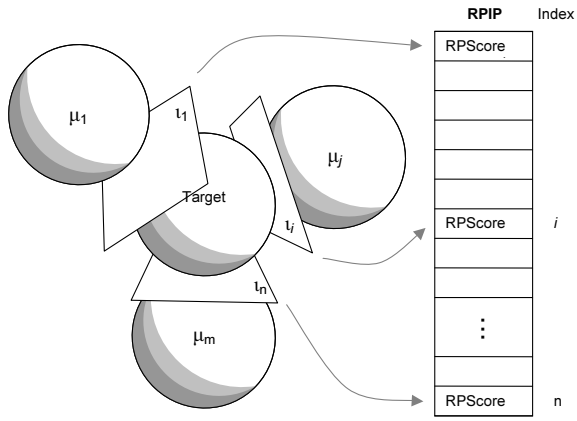

Figure 3
Construction of an RPIP. Each sphere represents the model (μ) of a death domain superfamily member and each plane represents an informative interface (ι). Note that every interface has an associated model, but that the number of models (m) may be different than the number of interfaces (n). Here, ι_1, ι_i and ι_n are associated with μ_1, μ_j and μ_m, respectively. The target's RPIP is the collection of RPScores calculated for each interface ι_i, for all $i \in [1...n]$.

In transforming the putative complexes into interaction profiles, we had to answer the question: If a pair of models forms 2000 putative complexes, how can we identify those that are most valuable for identifying family membership? Ideally, the algorithm should agree with the two known 3D protein complexes (APAF1/CASP9 and Pelle/Tube) that have been solved in the superfamily. Furthermore, we considered that a protein's interaction signature should not only consist of its binding domains and specific binding partners, but should be characterized by a gradient of different binding specificity. Our first intuition was to mine for *frequently occurring* putative complexes. To visualize what we mean by this, it is possible to plot all putative complexes in a 3D environment (see Figure 4(a)). The figure shows that points represent docking hits form clusters, and that these clusters represent frequently occurring putative complexes. However, contrary to our intuition, when these clusters were compared to the two known 3D protein complexes of death domain there were no spatial proximity between the position of the model with the position of the cluster representing the different docking solutions (Figure 4(b)). These data shows that the rescoring of the protein complexes solutions obtained by docking needed to be performed to eliminate false solutions from the real one. Effectively, GRAMM's method considers geometrical fit and hydrophobicity, but does not have optimal molecular mechanics force fields and desolvation parameters that can eliminate false positives. As a result we applied a rescoring algorithm [37] to the "raw" docking results to improve the ranking. The rescoring algorithm has the luxury of considering a much broader spectrum of factors due to the reduced search space. As a result, the list of complexes was re-ranked and a cutoff imposed with increased confidence, resulting in 11,260 complexes.

As a means of further filtering the docking results, we applied the notion of *sister list intersection*. Recall that sister lists are those that are returned from docking model μ_i against model μ_j and vice versa, for $i, j \in [1...m]$. The motivation behind sister list intersection is that, since we are only interested in domain-domain interactions, theoretically there should be no difference between the two lists. With this and the fact that the lists are predominantly false

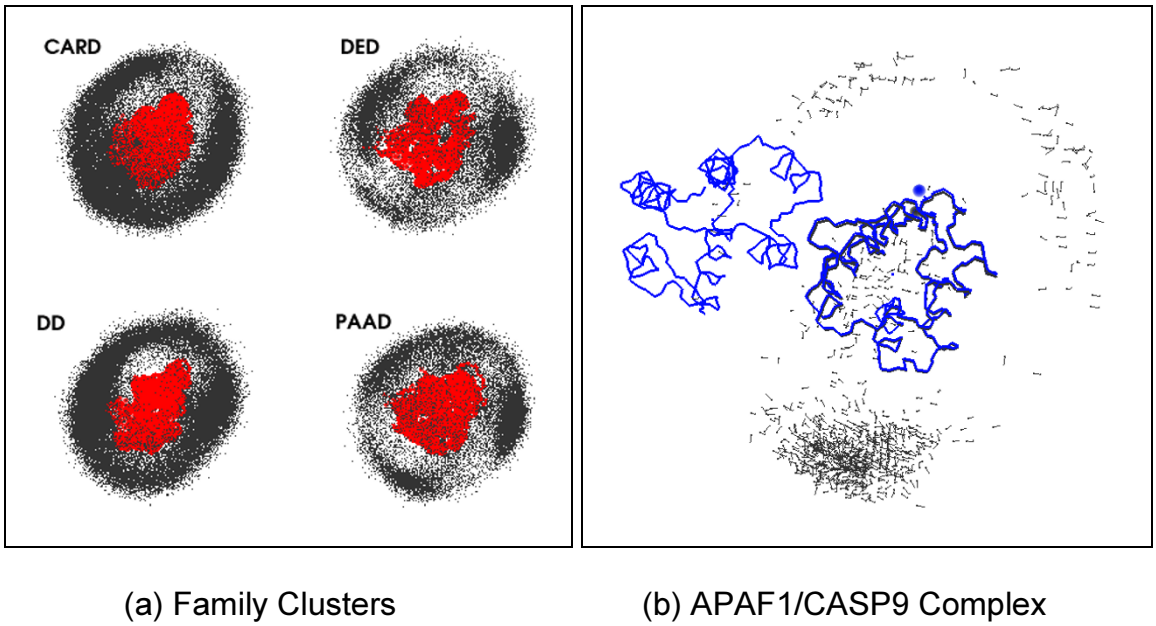

Figure 4
Unfiltered Predicted complexes. (a) Clusters of predicted complexes for each of the four families. Receptor models are shown as α-carbon traces in red. Each gray point indicates the center position of a GRAMM hit. (b) Predicted complexes *vs.* the known complex for APAF1/CASP9. The receptor (APAF1) of the predicted complexes is shown as α-carbon trace in gray; the ligand center points are shown as gray points with direction vectors. The known complex is shown in blue. Note that the known (blue) and predicted (gray) APAF1's have been aligned, and that the known CASP9 center is not within the dense regions of predicted centers.

positives in mind, we argue that sister list intersection is a logical means of filtering. When implementing the algorithm, we discovered that the routine for comparing complexes was best left with a level of flexibility. In other words, complexes need not be identical to be considered on the list, they simply must be sufficiently similar (recall Figure 2). The reason for this is that if identity were to be upheld, the resulting lists would not contain enough elements to allow for classification. By combining sister list intersection then rescoring the protein complex solution obtained from the raw data we were able to reduce the 1.1 million putative complexes down to 913 informative interfaces (i.e. n = 913). When we applied this mining procedure to the solved complex APAF1/CASP9, of the 2000 putative complexes returned from GRAMM output only two informative interfaces were found, one of which was the true interface. By filtering out the majority of putative complexes we were given the advantage of not having to perform exhaustive docking for each of the new family members we wanted to classify. Instead, we limited the translation/rotation space to the informative interfaces with the understanding that they form a representative population of the important interactions.

When arriving at the definition of the RPIP, the goal was to find a profile that was fast and easy to calculate, and that suitably described a protein's interaction behavior so that it could be successfully classified. We should note that the RPIP was not designed to have a direct physicochemical interpretation. As mentioned in the methods section, every RPIP element is associated with an informative interface, which in turn is associated with a model. Since every model corresponds to a member of the superfamily, we can assign a family to each of the RPIP elements. In this way, the RPIP can be divided into four sections, one for each family. In the family-based classifier we hypothesized that RPIP elements corresponding to the target's true family would, on average, be greater than the elements corresponding to the incorrect families. Figure 5 illustrates the hypothesized RPIP against a typical observed RPIP. As the figure implies, the family-based classifier performed poorly with an average 5-fold cross-validation accuracy of 61%. Alternative summary statistics for representing the families including mean, maximum and standard deviation, did not significantly improve the accuracy. Moreover, removing RPIP elements whose scores were consistently high or low across families simi-

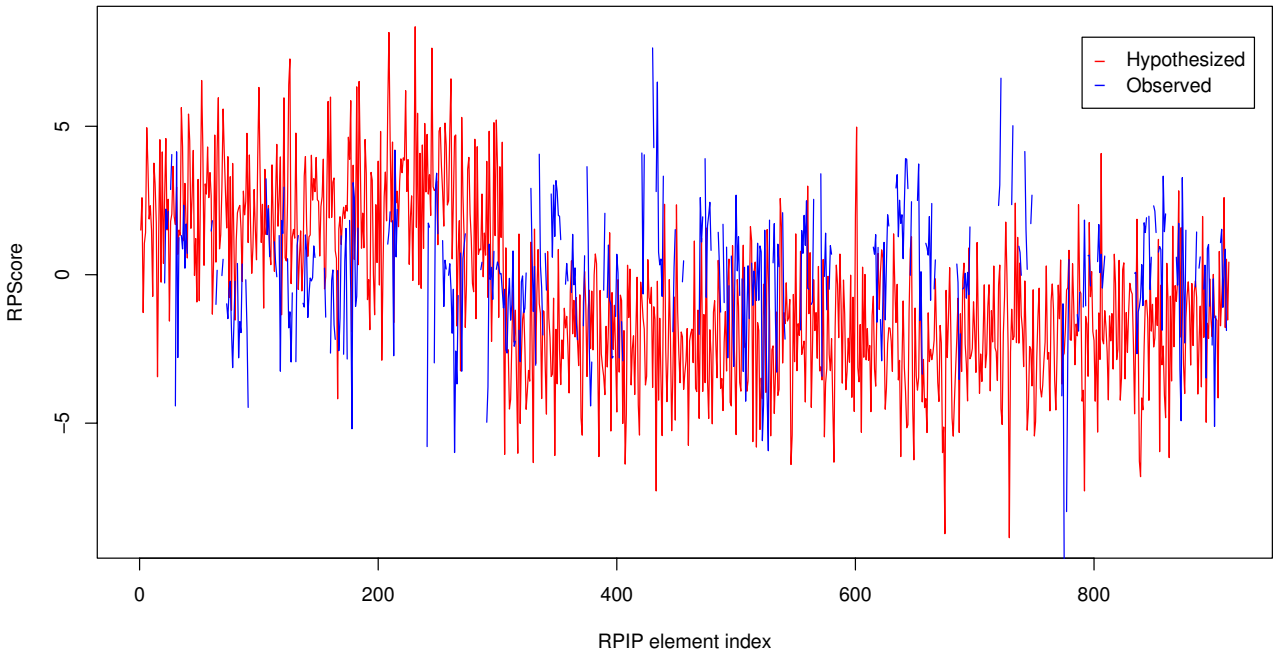

Figure 5
RPIP and family based classifier. This RPIP was generated from a CARD family member. The hypothesis of the family based classifier stated that RPIP elements corresponding to the target's true family should on average be greater than the others. The series in gray illustrates the predicted model where the RPScores corresponding to the CARD family (indices 0 through 304) should be generally greater than the rest. Note, however, that the observed values in black do not agree with the hypothesis, instead showing a generally even trend across all families.

larly did not result in higher accuracy (data not shown). The most obvious implication of these results is that an interface that is particularly stereo-chemically favorable to one family member is not necessarily favorable to another family member.

Unlike the family based classifier, the SVM-based classifier made no assumptions about how true family membership is related to RPIP element family membership; instead it relied on machine learning to automatically find correlations in the elements. The SVM-based classifier performed significantly better with an average cross-validation accuracy rate of 89%. Figure 6 shows an example SVM output from HODOCO. The increased performance of the SVM can be attributed to its ability to detect weaker patterns in the RPIPs, than the heuristic approach taken by the family based classifier. In particular, it can consider multiple simultaneous interaction propensities to arrive at its final decision. It is interesting to note that some sequences were consistently misclassified by the SVM. Figure 6 shows that APAF1's true family (CARD) is predicted to be the least probable amongst the four families. In this case, misclassification is due to the fact that it effectively has very unique binding and sequence properties compared to other members of the superfamily. It forms a supra-molecular complex called the apoptosome with CASPASE-9 (CASP9) and NAC [38] which involves a protein-protein interface not present in other family members.

Conclusions

The goal of this work was to show that *in silico* interaction-based protein classification can be obtained reliably for the death domain superfamily. We developed a classification pipeline that allows us to obtain protein classification.

While others have used *in silico* interaction profiles to characterize the docking ability of small molecules binding to experimentally determined 3D structures [39], or to discover novel protein interactions using known 3D complexes [40], our method is unique in that it applies to 3D molecular models of proteins complexes, and it not

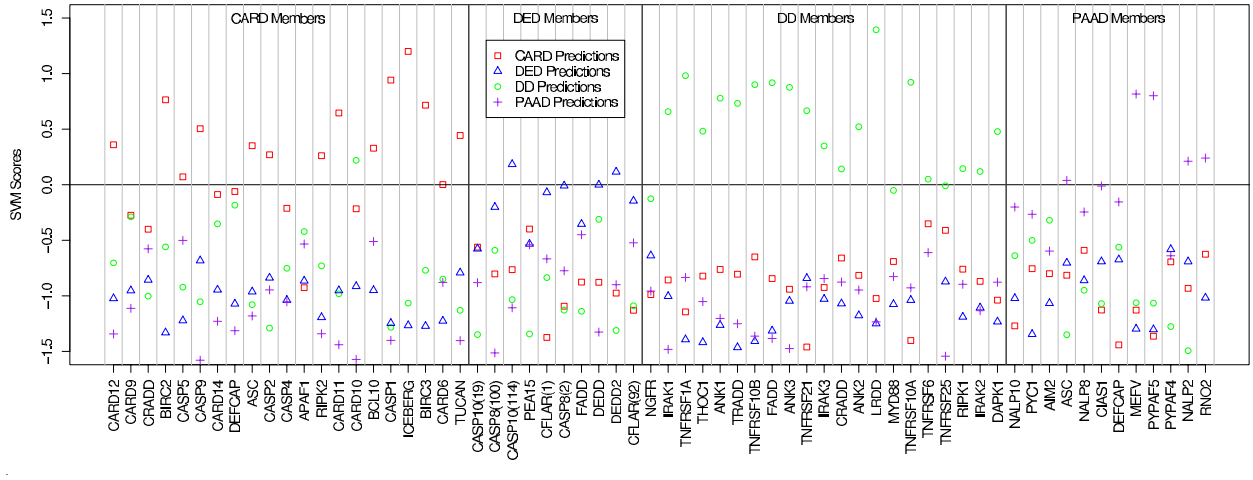

Figure 6
Classification of death domain family members using SVM. Each RPIP (*x*-axis) has four SVM scores associated with it; one from each family. The family with the highest score is the predicted family. Here, correct classifications include 19/21 CARD members; 8/10 DED members, with one near-tie; 21/21 DD members; and 10/12 PAAD members.

only considers multiple binding partners, but also multiple interfaces for each partner.

In future, there is much further work that can be done. Similar to [39], RPIPs could be used to cluster models rather than classify them. Here the goal would be to find alternative "families" based on interaction data, without regard to sequence homology. For example, discovering groups of models, sharing a common interaction interface or that are an outlier with a unique interface. Such a property has been highlighted by the constant misclassification of the CARD domain of APAF1 obtained in this study that has unique binding properties in the family.

It is very important to note that this study is only as powerful as the methods it builds upon. In particular, docking is still a highly active area of research with much work remaining to be done on model flexibility, solvent simulation, and force field optimization. Similarly, model building of the protein complex remains a difficult procedure requiring great care, making it difficult to accurately automate [41]. Our results have shown that despite the introduction of error in the classification pipeline due to the reliability of the underlying tools, interaction-profile based protein classification can be obtained with confidence. The fact that multiple parties have independently begun researching the potential of interaction profiles suggests that it may become a popular method for biological data mining in the future.

Methods

HODOCO is an *in silico* protein characterization and classification system consisting of three parts, a model building subsystem, a classifier building subsystem, and the classifier itself. An overview of the three major components is shown in Figure 1. When building the classifier the input to the system is a set of 3D structures, termed models (μ_1 through μ_m), that belong to the superfamily of interest. Since 3D structures for the majority of proteins are unsolved, a major aspect of the system is the model building subsystem (Figure 1(a)). Once the models are built, docking is performed to find putative complexes, which are then used to find what we call informative interfaces (ι_1 through ι_n to be discussed later). The informative interfaces are directly related to calculating the RPIPs, which in turn are used to build the classifier (Figure 1(b)). The model building subsystem and the classifier building subsystem have been designed to be fully independent so that improvements to their respective underlying tools can be applied to one without necessitating changes to the other. The final subsystem, the classifier, however is tightly bound to the second subsystem, and its distinction is mostly conceptual (Figure 1(c)). Each step in the system is now explained in detail.

Superfamily dataset collection

Three parallel approaches were used to obtain the set of human death domain protein sequences examined in this study. First, a literature survey was conducted to identify

an incomplete set of protein sequences from well-known family members. Second, the Pfam database was consulted to extract a set of protein sequences from members of the CARD, DED, DEATH and PAAD/DAPIN/PYRIN families. Third, the sequences found in the previous two methods were pooled to conduct a distant homologue search via an iterative BLAST procedure [25].

3D-structure modeling and alignment

Atomic coordinates from solved 3D protein structures were retrieved from the Protein Data Bank [26] and used for docking studies where available. Atomic coordinates from the remaining family members were obtained by homology modeling. Each amino acid sequence (the target sequence) from the death domain superfamily was submitted into the Polish metaserver [27]. Once Pair-wise alignments between the target sequence and a template were generated, a pair-wise alignment with the best score and the template's structure were submitted into the MODELLER program [28] to generate homology models for the target sequence. The default parameters for MODELLER were used, while the "loop-modelling" option was enabled. A total of 6 models were generated for each target, and the models were refined by molecular dynamics with simulated annealing (a functionality in MODELLER) to improve the quality of the model. All 6 models were verified for favorable geometrical and stereochemical properties using Verify 3D [29] and PROCHECK [30], and the rms deviation between the model and the template from which the model originated. From these criteria the best one was selected as the representative model for the target. Low quality models were discarded if they exhibited an RMSD greater than 1.5 Å on the total main chain atoms with the structure template from which they were built. All remaining models were then used as input to the classification analysis, as if they were the original input to the system. Models were structurally superimposed on a reference model (ASC) using the Combinatorial Extension method [31] for docking studies. We refer to the models as μ_1 through μ_m, where m is the number of models.

Model interaction prediction

Putative complexes were predicted through computational docking using GRAMM [32]. GRAMM performs an exhaustive 6-dimensional search of all translations and rotations between a given pair of macromolecules and returns a list of high scoring complexes based on rigid-body geometric fit and hydrophobicity. It should be mentioned that, due to GRAMM's algorithm, the order of the models is important. Specifically, the list of complexes returned from docking molecule *A* against molecule *B* is not guaranteed to be the same as the list returned from docking molecule *B* against molecule A. We refer to these related lists as *sister lists*. GRAMM's parameters chosen such that the two known death domain superfamily complexes, human caspase recruitment domains of APAF1 with pro-Caspase-9 (PDB: 3YGS) and drosophila death domains of PELLE with TUBE (PDB: 1D2Z) had the best rank possible when docking their respective individual monomers were: Matching mode = generic, grid step = 1.5, repulsion attraction = 5, attraction double range = 0, potential range type = atom_radius, projection = black and white, representation = all, number of matches to output = 1000 and angle of rotation = 10. Each model within each family was docked against every other model within the same family and the top 1000 complexes from each docking simulation were retained in a MySQL database for further filtering. It could be argued that the GRAMM parameters giving the best rank, when obtained from the docking of the individual monomers of the two known protein complexes (bound form), could not be optimal when each death domain superfamily member are docked against each others (unbound form). Effectively docking from unbound monomers has to consider conformational changes of the protein partners upon binding that do not occur when a protein monomer belongs to a known protein complex. To limit the complexity of the problem and allow comparison between our simulations we used the low resolution docking parameters of GRAMM. In such a procedure flexibility is handled through a "soft belt" into which atoms from the second molecule can penetrate reducing drastically the complexity [22], and increasing the speed of the simulation.

Mining for informative interfaces

The database of resulting complexes was mined for those with the maximum information gain with respect to family classification. We refer to the mined complexes as *informative interfaces*, since each complex defines an interface between the component models. We labeled the informative interfaces ι_1 through ι_n, n the number of interfaces. The mining procedure consisted of two parts: (i) Rescoring the complexes and (ii) Taking the intersection of sister lists, described shortly. All mining was performed via a combination of shell scripts, Perl scripts and C++ programs.

Rescoring of complexes

The first mining technique was to rescore each complex using the software Rpdock – a member of the 3D-Dock suite [33] and reject those below a threshold. Rpdock uses evidence gathered empirically to quantify the probability of a complex's existence and returns a score (RPScore) based on the results. The algorithm uses residue pair potentials across protein interfaces [34] as the basis for the score. Ranked via this alternative scoring system, the top 10 complexes from each docking experiment were retained, while the other 990 complexes were rejected.

Intersection of sister lists

The second mining technique applied took advantage of the input-model order dependence of GRAMM. Every pair of sister lists was examined for complexes found in both lists. More precisely, every pair-wise combination of complexes from all sister lists was considered in turn. Recall that complexes in sister lists are composed of the same two models (call them μ_i and μ_j, $i, j \in [1...m]$). If after the two instances of μ_i were structurally superimposed, the distance between the two instances of μ_j fell below a given threshold then both complexes were retained. Complexes never retained in this way were rejected (see Figure 2). Note that the two instances of μ_j need not be exactly structurally superimposed. We refer to μ_i as the *aligned model* and μ_j as the *orientation-independent model*. The informative interface is then the interface that lies between these models.

RPIP construction

The set of informative interfaces was used to generate vectors of RPScores; one vector per model. Each RPScore was calculated as follows.

1. Given the 3D structure of a sequence of interest (the target), and an informative interface, ι_i ($i \in [1...n]$), consisting of an aligned model and an orientation independent model, the aligned model was replaced with the target ensuring preservation of orientation. (Recall that all models were previously structurally superimposed to a reference model, thereby normalizing rotations across models). We refer to this modified complex as the *hybrid complex*.

2. Rpdock was used to calculate the RPScore of the hybrid complex, and the result was stored in element i of the target's RPIP. Note that the RPScore of the hybrid complex could be grossly different from that of the unmodified complex.

For a given model, the collection of RPScores resulting from performing the above calculation on each of the informative interfaces was placed into a vector and termed the model's Residue Potential Interaction Profile, or RPIP (see Figure 3). Note that there is a one-to-one correspondence between RPIP elements and informative interfaces. To avoid bias, all informative interfaces that involved the model itself were assigned a null RPScore and ignored in all future steps.

Building the classifier

Two methods of classification were attempted. First, it was postulated that the RPIP elements pertaining to a model's true family would be, on average, greater than those not. Thus, a classifier was built that compared the median RPScores across RPIP elements for each of the four families, and made a prediction based on the greatest mean. This classifier was termed the *family-based classifier*.

The second method used the RPIPs to build four support vector machines one for each family. This classifier was termed the *SVM-based classifier*. The software SVM light was used in this study http://svmlight.joachims.org. Each machine was trained to discriminate members of one family, so that all four machines would have to be used to make a final prediction. A linear kernel was used when building the machines to avoid undue distortion of the underlying RPScores.

Testing

To test the accuracy of HODOCO we conducted five runs of 5-fold cross-validation for each classification method. Generally, k-fold cross-validation randomly divides the target dataset into k equal-sized partitions, iteratively using one partition for testing and the other partitions for training. Method accuracy is measured as the average number of correctly classified models divided by the total number of models over a series of cross-validation runs. Referring back to Figure 1(c), a target is classified by building its model, building its RPIP and finally applying one of the classification methods to the RPIP. Note that it is assumed that the unknown sequence has been previously screened to be a member of the superfamily.

Authors' contributions

MS performed the first docking experiments of the project and developed clustering methods to classify protein based on protein-protein interactions, DL contributed to the writing of the manuscript and to data analysis by implementing the support vector machine. FP performed part of the docking experiments, design the necessary experiments to complete this work, provided guidance to mine the data and contributed to the writing of the manuscript.

Acknowledgments

We would especially like to thank Maggie Lau and Tony Zhan for performing much of the pre-analytical data gathering, and the SFU Co-operative Education Program for making their employment possible. D.L.'s stipend was provided by the Canadian Institute for Health Research (CIHR) and by the Michael Smith Foundation. We are also grateful for funding support from the Natural Sciences and Engineering Research Council of Canada (NSERC) and by the British Columbia Advanced Systems Institute (BC ASI).

References

1. Ponstingl M, Henrick K, Thornton JM: **Discriminating between homodimeric and monomeric proteins in the crystalline state.** *Proteins* 2000, **41(1):**47-57.
2. Valdar WS, Thornton JM: **Conservation helps to identify biologically relevant crystal contacts.** *J Mol Biol* 2001, **313(2):**399-416.
3. Mintseris J, Weng Z: **Atomic Contact Vector in Protein-Protein Recognition.** *Proteins* 2003, **53:**629-639.
4. Wu CH, Nikolskaya A, Huang H, Yeh LS, Natale DA, Vinayaka CR, Hu ZZ, Mazumder R, Kumar S, Kourtesis P, Ledley RS, Suzek BE, Arminski L, Chen Y, Zhang J, Cardenas JL, Chung S, Castro-Alvear J,

Dinkov G, Barker WC: **PIRSF: family classification system at the Protein Information Resource.** *Nucleic Acids Res* **2004:**D112-114. 32 Database issue
5. Ouzounis CA, Coulson RM, Enright AJ, Kunin V, Pereira-Leal JB: **Classification schemes for protein structure and function.** *Nat Rev Genet* 2003, **4(7):**508-19.
6. Sigrist CJ, Cerutti L, Hulo N, Gattiker A, Falquet L, Pagni M, Bairoch A, Bucher P: **PROSITE: a documented database using patterns and profiles as motif descriptors.** *Brief Bioinform* 2002, **3(3):**265-274.
7. Attwood TK, Beck ME, Flower DR, Scordis P, Selley JN: **The PRINTS protein fingerprint database in its fifth year.** *Nucleic Acids Res* 1998, **26(1):**304-308.
8. Schultz J, Milpetz F, Bork P, Ponting CP: **SMART, a simple modular architecture research tool: identification of signaling domains.** *Proc Natl Acad Sci U S A* 1998, **95(11):**5857-5864.
9. Servant F, Bru C, Carrere S, Courcelle E, Gouzy J, Peyruc D, Kahn D: **ProDom: automated clustering of homologous domains.** *Brief Bioinform* 2002, **3(3):**246-251.
10. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna , Marshall M, Moxon S, Sonnhammer EL, *et al.*: **The Pfam protein families database.** *Nucleic Acids Res* 2004, **32:**D138-141. 32 Database issue
11. Murzin AG, Brenner SE, Hubbard T, Chothia C: **SCOP: a structural classification of proteins database for the investigation of sequences and structures.** *J Mol Biol* 1995, **247(4):**536-540.
12. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM: **CATH – a hierarchic classification of protein domain structures.** *Structure* 1997, **5(8):**1093-1108.
13. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Barrell D, Bateman A, Binns D, Biswas M, Bradley P, Bork P, *et al.*: **The InterPro Database, 2003 brings increased coverage and new features.** *Nucleic Acids Res* 2003, **31(1):**315-318.
14. Huang H, Barker WC, Chen Y, Wu CH: **iProClass: an integrated database of protein family, function and structure information.** *Nucleic Acids Res* 2003, **31(1):**390-392.
15. von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P: **Comparative assessment of large-scale data sets of protein-protein interactions.** *Nature* 2002, **417(6887):**399-403.
16. Vazquez A, Flammini A, Maritan A, Vespignani A: **Global protein function prediction from protein-protein interaction networks.** *Nat Biotechnol* 2003, **21(6):**697-700.
17. Li S, Armstrong CM, Bertin N, Ge H, Milstein S, Boxem M, Vidalain PO, Han JD, Chesneau A, Hao T, Goldberg DS, Li N, Martinez M, Rual JF, Lamesch P, Xu L, Tewari M, Wong SL, Zhang LV, Berriz GF, Jacotot L, Vaglio P, Reboul J, Hirozane-Kishikawa T, Li Q, Gabel HW, Elewa A, Baumgartner B, Rose DJ, Yu H, Bosak S, Sequerra R, Fraser A, Mango SE, Saxton WM, Strome S, Van Den Heuvel S, Piano F, Vandenhaute J, Sardet C, Gerstein M, Doucette-Stamm L, Gunsalus KC, Harper JW, Cusick ME, Roth FP, Hill DE, Vidal M: **A Map of the Interactome Network of the Metazoan C. elegans.** *Science* 2004, **303(5657):**540-543.
18. Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E, Vijayadamodar G, Pochart P, Machineni H, Welsh M, Kong Y, Zerhusen B, Malcolm R, Varrone Z, Collis A, Minto M, Burgess S, McDaniel L, Stimpson E, Spriggs F, Williams J, Neurath K, Ioime N, Agee M, Voss E, Furtak K, Renzulli R, Aanensen N, Carrolla S, Bickelhaupt E, Lazovatsky Y, DaSilva A, Zhong J, Stanyon CA, Finley RL Jr, White KP, Braverman M, Jarvie T, Gold S, Leach M, Knight J, Shimkets RA, McKenna MP, Chant J, Rothberg JM: **A protein interaction map of Drosophila melanogaster.** *Science* 2003, **302(5651):**1727-1736.
19. Strong M, Graeber TG, Beeby M, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D: **Visualization and interpretation of protein networks in Mycobacterium tuberculosis based on hierarchical clustering of genome-wide functional linkage maps.** *Nucleic Acids Res* 2003, **31(24):**7099-7109.
20. Janin J, Henrick K, Moult J, Eyck LT, Sternberg MJ, Vajda S, Vakser I, Wodak SJ: **Critical Assessment of PRedicted Interactions. CAPRI: a Critical Assessment of PRedicted Interactions.** *Proteins* 2003, **52(1):**2-9.
21. Venclovas C, Zemla A, Fidelis K, Moult J: **Assessment of progress over the CASP experiments.** *Proteins* 2003, **53(6):**585-95.
22. Halperin I, Buyong M, Wolfson H, Nussinov R: **Principle of Docking: An Overview of Search Algorithms and a guide to Scoring Functions.** *Proteins* 2002, **47:**409-443.
23. Moult J, Fidelis K, Zemla A, Hubbard T: **Critical assessment of methods of protein structure prediction (CASP)-round V.** *Proteins* 2003, **53:**334-339.
24. Smith GR, Sternberg MJ: **Prediction of protein-protein interactions by docking methods.** *Curr Opin Struct Biol* 2002, **12(1):**28-35.
25. Li W, Pio F, Pawlowski K, Godzik A: **Saturated BLAST: an automated multiple intermediate sequence search used to detect distant homology.** *Bioinformatics* 2000, **16(12):**1105-1110.
26. Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, Feng Z, Gilliland GL, Iype L, Jain S, *et al.*: **The Protein Data Bank.** *Acta Crystallogr D Biol Crystallogr* 2002, **58(Pt 6 No 1):**899-907.
27. Bujnicki JM, Elofsson A, Fischer D, Rychlewski L: **Structure prediction meta server.** *Bioinformatics* 2001, **17(8):**750-751.
28. Sali A, Blundell TL: **Comparative protein modelling by satisfaction of spatial restraints.** *J Mol Biol* 1993, **234(3):**779-815.
29. Luthy R, Bowie JU, Eisenberg D: **Assessment of protein models with three-dimensional profiles.** *Nature* 1992, **356(6364):**83-85.
30. Laskowski RA: **PDBsum: summaries and analyses of PDB structures.** *Nucleic Acids Res* 2001, **29(1):**221-222.
31. Shindyalov IN, Bourne PE: **Protein structure alignment by incremental combinatorial extension (CE) of the optimal path.** *Protein Eng* 1998, **11(9):**739-747.
32. Katchalski-Katzir E, Shariv I, Eisenstein M, Friesem AA, Aflalo C, Vakser IA: **Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques.** *Proc Natl Acad Sci U S A* 1992, **89(6):**2195-2199.
33. Smith GR, Sternberg MJ: **Evaluation of the 3D-Dock protein docking suite in rounds 1 and 2 of the CAPRI blind trial.** *Proteins* 2003, **52(1):**74-79.
34. Moont G, Gabb HA, Sternberg MJ: **Use of pair potentials across protein interfaces in screening predicted docked complexes.** *Proteins* 1999, **35(3):**364-373.
35. Reed JC, Doctor K, Rojas A, Zapata JM, Stehlik C, Fiorentino L, Damiano J, Roth W, Matsuzawa S, Newman R, *et al.*: **Comparative analysis of apoptosis and inflammation genes of mice and humans.** *Genome Res* 2003, **13(6B):**1376-1388.
36. Pawlowski K, Godzik A: **Surface map comparison: studying function diversity of homologous proteins.** *J Mol Biol* 2001, **309(3):**793-806.
37. Smith GR, Sternberg MJ: **Prediction of protein-protein interactions by docking methods.** *Curr Opin Struct Biol* 2002, **12(1):**28-35.
38. Chu ZL, Pio F, Xie Z, Welsh K, Krajewska M, Krajewski S, Godzik A, Reed JC: **A novel enhancer of the Apaf1 apoptosome involved in cytochrome c-dependent caspase activation and apoptosis.** *J Biol Chem* 2001, **276(12):**9239-9245.
39. Hayashi Y, Sakaguchi K, Kobayashi M, Kikuchi Y, Ichiishi E: **Molecular evaluation using in silico protein interaction profiles.** *Bioinformatics* 2003, **19(12):**1514-1523.
40. Aloy P, Russell RB: **Interrogating protein interaction networks through structural biology.** *Proc Natl Acad Sci USA* 2002, **99(9):**5896-5901.
41. Eyrich VA, Przybylski D, Koh IY, Grana O, Pazos F, Valencia A, Rost B: **CAFASP3 in the spotlight of EVA.** *Proteins* 2003, **53(Suppl 6):**548-560.
42. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D: **The human genome browser at UCSC.** *Genome Res* 2002, **12(6):**996-1006.