Final Progress Report for Research Projects Funded by Health Research Grants

Instructions: Please complete all of the items as instructed. Do not delete instructions. Do not leave any items blank; responses must be provided for all items. If your response to an item is "None", please specify "None" as your response. "Not applicable" is not an acceptable response for any of the items. There is no limit to the length of your response to any question. Responses should be single-spaced, no smaller than 12-point type. The report **must be completed using MS Word**. Submitted reports must be Word documents; they should not be converted to pdf format. Questions? Contact Health Research Program staff at 717-783-2548.

1. **Grantee Institution:** The Institute for Cancer Research

2. **Reporting Period (start and end date of grant award period):** 1/1/2011 – 6/30/2014

3. **Grant Contact Person (First Name, M.I., Last Name, Degrees):** Maria Minko Gill

4. **Grant Contact Person's Telephone Number:** 215-728-2659

5. **Grant SAP Number:** 4100054848

6. **Project Number and Title of Research Project:** 05- Classification and Prediction of Protein-Protein Interactions in Biology and Medicine

7. **Start and End Date of Research Project:** 1/1/2011 – 6/30/2013

8. **Name of Principal Investigator for the Research Project:** Roland Dunbrack, Ph.D.

9. **Research Project Expenses**.

   9(A) Please provide the total amount of health research grant funds spent on this project for the entire duration of the grant, including indirect costs and any interest earned that was spent:

   $ 922,057

   9(B) Provide the last names (include first initial if multiple individuals with the same last name are listed) of **all** persons who worked on this research project and were supported with health research funds. Include position titles (Principal Investigator, Graduate Assistant, Post-doctoral Fellow, etc.), percent of effort on project and total health research funds expended for the position. For multiple year projects, if percent of effort varied from year to year, report in the % of Effort column the effort by year 1, 2, 3, etc. of the project (x% Yr 1; z% Yr 2-3).

Last Name, First Name	Position Title	% of Effort on Project	Cost
Dunbrack, Roland L	PI	20% year 1; 11% year 2; 2% year 3	$56,455.40
Lehmann, Andreas	Postdoctoral Associate	54% year 2	$40,791.66
North, Benjamin H	Postdoctoral Associate	54% year 1	$37,309.30
Shapovalov, Maxim V	Senior Program Analyst	50% year 1; 100% year 2	$103,400.07
Wei, Qiong	Research Associate	100% years 1-2; 10% year 3	$91,244.68
Xu, Qifang	Programmer Analyst II	100% years 1-2	$143,428.19

9(C) Provide the names of **all** persons who worked on this research project, but who *were not* supported with health research funds. Include position titles (Research Assistant, Administrative Assistant, etc.) and percent of effort on project. For multiple year projects, if percent of effort varied from year to year, report in the % of Effort column the effort by year 1, 2, 3, etc. of the project (x% Yr 1; z% Yr 2-3).

Last Name, First Name	Position Title	% of Effort on Project
None		

9(D) Provide a list of **all** scientific equipment purchased as part of this research grant, a short description of the value (benefit) derived by the institution from this equipment, and the cost of the equipment.

Type of Scientific Equipment	Value Derived	Cost
None		

10. Co-funding of Research Project during Health Research Grant Award Period. Did this research project receive funding from any other source during the project period when it was supported by the health research grant?

Yes_____ No___X_____

If yes, please indicate the source and amount of other funds:

11. Leveraging of Additional Funds

11(A) As a result of the health research funds provided for this research project, were you able to apply for and/or obtain funding from other sources to continue or expand the research?

Yes____X_____ No_____

If yes, please list the applications submitted (column A), the funding agency (National Institutes of Health—NIH, or other source in column B), the month and year when the application was submitted (column C), and the amount of funds requested (column D). If you have received a notice that the grant will be funded, please indicate the amount of funds to be awarded (column E). If the grant was not funded, insert "not funded" in column E.

Do not include funding from your own institution or from CURE (tobacco settlement funds). Do not include grants submitted prior to the start date of the grant as shown in Question 2. If you list grants submitted within 1-6 months of the start date of this grant, add a statement below the table indicating how the data/results from this project were used to secure that grant.

A. Title of research project on grant application	B. Funding agency (check those that apply)	C. Month and Year Submitted	D. Amount of funds requested:	E. Amount of funds awarded:
R01 - Bayesian statistics and algorithms for homology modeling	☒NIH ☐ Other federal (specify:_____) ☐ Nonfederal source (specify:_)	July 2012	$1,785,000	$1,356,000
R01 - Dimensionality Reduction of Protein Structural Ensembles	☒ NIH ☐ Other federal (specify:_____) ☐ Nonfederal source (specify:_)	October 2013	$2,212,853	Not funded

11(B) Are you planning to apply for additional funding in the future to continue or expand the research?

Yes____X_____ No_____

If yes, please describe your plans:

We are planning to write an NIH grant on kinase structural bioinformatics, which will use data from this project and this will require maintaining and developing the existing database and webserver. We may also apply for a software/database/webserver maintenance grant from NIH.

12. Future of Research Project. What are the future plans for this research project?

We have several plans for the database and webserver. First, there are many potential hypotheses that can be generated from the large clusters that are mostly not annotated by crystallographers. Many of these may contain biologically relevant interactions that have not been previously recognized. These will be examined over time and potential research

projects with collaborators initiated. Second, we would like to perform the analysis on the clan level. This has had to be postponed because of the very large amount of CPU required and the difficulty of aligning very distantly related proteins. But it is still of interest in the long run. Third, while we have identified proteins that interact with other proteins in the way that peptides do (e.g. full protein domains that bind to PDZ domains in the PDB; the autophosphorylation structures, etc.), we have not analyzed these in detail because it takes time to investigate the relevant literature on each domain family to determine if the interactions are biologically relevant. This will be an ongoing project using ProtCID. Finally, Pfam is a limitation in some cases, and some form of automatic clustering of domains and structures covering domains not found in Pfam or for which Pfam has HMMs covering multiple domains may be required.

13. **New Investigator Training and Development**. Did students participate in project supported internships or graduate or post-graduate training for at least one semester or one summer?

    Yes____X_____ No_____

    If yes, how many students? Please specify in the tables below:

	Undergraduate	Masters	Pre-doc	Post-doc
Male				2
Female				1
Unknown				
Total				**3**

	Undergraduate	Masters	Pre-doc	Post-doc
Hispanic				
Non-Hispanic				3
Unknown				
Total				**3**

	Undergraduate	Masters	Pre-doc	Post-doc
White				2
Black				
Asian				1
Other				
Unknown				
Total				**3**

14. **Recruitment of Out-of–State Researchers**. Did you bring researchers into Pennsylvania to carry out this research project?

    Yes_____ No_____X_____

If yes, please list the name and degree of each researcher and his/her previous affiliation:

15. Impact on Research Capacity and Quality. Did the health research project enhance the quality and/or capacity of research at your institution?

Yes____X____ No_____

If yes, describe how improvements in infrastructure, the addition of new investigators, and other resources have led to more and better research.

The database and website created under this research project have been used extensively in the structure prediction of protein complexes in the Molecular Modeling Facility at Fox Chase Cancer Center. This Facility serves research groups at Fox Chase and at Temple University School of Medicine, of which we are a part.

16. Collaboration, business and community involvement.

16(A) Did the health research funds lead to collaboration with research partners outside of your institution (e.g., entire university, entire hospital system)?

Yes____X_____ No_____

If yes, please describe the collaborations:

The database was used to create benchmarking data sets for 4 different research groups who study protein-protein interactions and protein-peptide interactions. These research groups were at MIT, Hebrew University, Albert Einstein College of Medicine, and Yale University.

16(B) Did the research project result in commercial development of any research products?

Yes_____ No_____X_____

If yes, please describe commercial development activities that resulted from the research project:

16(C) Did the research lead to new involvement with the community?

Yes_____ No____X_____

If yes, please describe involvement with community groups that resulted from the research project:

17. Progress in Achieving Research Goals, Objectives and Aims.

List the project goals, objectives and specific aims (as contained in the grant agreement). Summarize the progress made in achieving these goals, objectives and aims for the period that the project was funded (i.e., from project start date through end date). Indicate whether or not each goal/objective/aim was achieved; if something was not achieved, note the reasons why. Describe the methods used. If changes were made to the research goals/objectives/aims, methods, design or timeline since the original grant application was submitted, please describe the changes. Provide detailed results of the project. Include evidence of the data that was generated and analyzed, and provide tables, graphs, and figures of the data. List published abstracts, poster presentations and scientific meeting presentations at the end of the summary of progress; peer-reviewed publications should be listed under item 20.

This response should be a DETAILED report of the methods and findings. It is not sufficient to state that the work was completed. Insufficient information may result in an unfavorable performance review, which may jeopardize future funding. If research findings are pending publication you must still include enough detail for the expert peer reviewers to evaluate the progress during the course of the project.

Health research grants funded under the Tobacco Settlement Act will be evaluated via a performance review by an expert panel of researchers and clinicians who will assess project work using this Final Progress Report, all project Annual Reports and the project's strategic plan. After the final performance review of each project is complete, approximately 12-16 months after the end of the grant, this Final Progress Report, as well as the Final Performance Review Report containing the comments of the expert review panel, and the grantee's written response to the Final Performance Review Report, will be posted on the CURE Web site.

There is no limit to the length of your response. Responses must be single-spaced below, no smaller than 12-point type. If you cut and paste text from a publication, be sure symbols print properly, e.g., the Greek symbol for alpha (α) and beta (ß) should not print as boxes (□) and include the appropriate citation(s). DO NOT DELETE THESE INSTRUCTIONS.

The main goals of the project were to predict the structures of multimeric protein complexes involved in important biological processes associated with the development and prevention of cancer and to provide a resource for other researchers to perform similar analyses. These goals were accomplished by exploiting structural information on proteins and protein complexes available in the Protein Data Bank (PDB). The Specific Aims were:

1. A complete and continuously up-to-date domain classification of all known protein structures.

a. Optimizing improved sequence-based methods for classifying protein structures.

b. Using structural information to identify distant evolutionary relationships and merge distantly related protein families into superfamilies or clans.

c. Development of an algorithm to assign the domain classification to multi-domain proteins. The algorithm must handle split domains and domain insertions properly.

d. Development of special techniques for assigning domain segments in repeat proteins.
2. Predicting the structures of protein-protein interactions from the PDB
a. Determining the interactions of domains from different families in all available protein structures.
b. Analysis of domain-domain interactions in very distantly related proteins.
c. Determining the evolutionarily conserved interactions of protein domains with nucleic acids, ions, peptides, and common ligands.
3. Studies of specific protein families involved in cell signaling events associated with cancer.
a. The structures of autophosphorylation events of protein kinases.
b. The role of homodimeric and protein-protein interactions in kinase activation and inhibition.

The results of these studies are as follows.

1. Pfam assignments to structures in the Protein Data Bank
In order to investigate the interactions of homologous domains in the PDB in different crystals, we required an ongoing classification of structures in the PDB into homologous groups. SCOP (Structural Classification of Proteins) and CATH (Class-Architecture-Topology-Homology) (and other databases are not up-to-date with respect to the PDB and require manual creation of new folds and superfamilies. Pfam, on the other hand, is a sequence-based clustering of protein domains, whether they are of known structure or not. Pfam sequence-based domains also have biological annotations and names that reflect biological function, such as SH2 (Src homology 2) and Pkinase. We therefore decided to develop an automatic method for assigning Pfams to the PDB, including potentially remote relationships of structures (and their sequences) to Pfam domains and the careful annotation of split domains – those that contain long insertions relative to the Pfam domain definitions. The database is called "PDBfam" and is located at http://dunbrack2.fccc.edu/protcid/pdbfam.

We approach this task in four steps in a process performed monthly. First, we use consensus sequences derived from PSI-BLAST profiles for each PDB sequence to search against the hidden Markov models (HMMs) in PFAM-A. These consensus sequences often provide better E-values and longer alignments than the original sequences. Second, we use HHBlits and HHSearch to align hidden Markov models for each PDB sequence to each Pfam HMM. These alignments are often longer and identify more remote relationships than sequence-HMM alignments. Third, we use structure alignments to compare weak hits or those with short alignments to structures with the highest statistical significance for each PFAM domain. These alignments are often of high statistical significance and are sometimes much longer than the sequence and HMM-based alignments. Fourth, we input these data to a greedy algorithm that accurately handles domains split by large insertions or that includes separate chains in a PDB structure. This algorithm is much more realistic and sensible than the one that PFAM itself uses. We examined the results of each of these in turn as determined during the development of the method. Numbers below reflect data in 2011-2013 at the time the comparisons were performed. The data come from various time points and therefore may cover smaller or larger number of PDB entries and Pfams.

Consensus sequences. We derived consensus sequences from PSI-BLAST profiles for every unique sequence in the PDB. One round of PSI-BLAST was used to search Uniprot with an E-value cutoff of 0.001 for sequences used to produce the profile. The "log-odds consensus

sequence" for each PDB sequence was derived by taking the highest scoring amino acid type at each position of the position-specific scoring matrix produced by PSI-BLAST. The "percentage consensus sequence" was derived by taking the most common amino acid at each position of the multiple sequence alignment produced by PSI-BLAST.

To investigate whether the consensus sequences improve the sensitivity of PFAM assignments by HMMER 3.0, we took the single best-scoring PFAM domain assignment for each PDB unique sequence from both consensus sequence types and the original sequences ("PDB-orig") with E-value < 0.001, then compared the distributions of E-values. Figure 1 shows the kernel density estimates and cumulative functions of $\log_{10}$(E-value). The consensus results (cyan and green) are shifted to significantly better E-values (more negative $\log_{10}$(E-value)) than the original PDB sequences (red). The purple curve shows the result of choosing the best E-value of all three sequences ("PDB-best"). The mode of the density for the original sequences is at about 1.0E-20 while for the consensus sequences it is about 1.0E-25.

For each unique sequence, we selected the PFAM assignments which are in the same PFAM and cover approximately the same sequence regions with E-value < 0.001, then produced scatter plots of the PDB-orig alignment lengths against the consensus sequences and the PDB-best result (Figure 2). About 80% of the consensus sequence alignments are longer than the original sequence alignments. In some cases, PFAM assignments of consensus sequences are broken down into two or more fragments, producing separate alignments that are shorter than the single alignment produced by the original sequence, producing points far below the diagonal. This situation is fixed in the algorithm that pieces together the separate PFAM assignments into an assignment for each whole sequence.

Table 1 shows the results of scoring each PFAM HMM against each unique PDB sequence and the consensus sequences. The number of PFAMs, the number of unique sequences with PFAM hits, etc. are shown in the table for those PFAMs with E-values better than 0.001. The consensus sequences improve both in the coverage of PDB sequences and the coverage of PFAM A models. The consensus sequences find 6.1% more PFAMs in the PDB than the original sequences, and cover 3.5% more residues of the unique PDB sequences.

<u>HMM-HMM alignments.</u> We have incorporated the use of HMM-HMM (hidden Markov model) alignments into PDBfam using the program HHSearch into PDBfam. These alignments are capable of finding much more remote relationships than the PSI-BLAST consensus sequence alignments. However, one has to be careful how these remote relationships are used. The problem is that HMM-HMM alignments often score more remote Pfam HMMs to a PDB HMM better than much closer HMMs easily assigned by HMMER sequence-HMM alignments. The goal is to assign a PDB sequence to the closest Pfam family, not more remote ones in the same superfamily (or "clan" as Pfam calls them). If applied to the whole PDB, HHSearch makes such mistakes for about 8% of sequences, which is substantial enough to require a way to deal with this problem. Our solution was to assign the original-sequence and the PSI-BLAST consensus-sequence HMMER alignments first up to the E-value cutoff we established for the greedy algorithm (see below) and then to utilize the HMM-HMM alignments. With the addition of HHSearch, we now have assignments of Pfams to 99% of PDB entries and 99.3% of unique sequences > 50 amino acids. The final results and comparison to other sources of domain

assignments to the PDB are shown in Table 2 (data as of 2012).

Structure alignment. We used structure alignment to verify PFAM assignments with poor E-values and/or alignments that do not fully cover the HMM. To do this, we chose "exemplars" for each PFAM found in the PDB, if the best E-value over all PDB sequences or consensus sequences was $\leq 10^{-5}$ and the alignment covered at least 80% of the HMM positions with coordinates from a structure. Of the 7,785 PFAMs we find in the PDB (with any E-value), 63% of them (4,884) have an exemplar under these definitions.

For each PFAM assignment with E-value $> 10^{-5}$ or missing more than 10 residues of the HMM on either end of the alignment, we checked whether there was an exemplar for that PFAM or any PFAM in the same clan. PFAM defines clans for two or more PFAMs that are homologous to each other. This is similar to superfamily definitions in such databases as SCOP. These PDB sequences with poor hits fall into 5,258 PFAMs, and comprise 20,685 (14.8%) of assignments to all PDB sequences. We used the FATCAT program from Adam Godzik's group to align the structures of proteins with poor PFAM assignments to the exemplars in the same clan as the original PFAM assignment. FATCAT was able to produce alignments with p-values better than 0.001 for 15,580 (75.3%) of these assignments. This is 11.1% of the total assignments to PDB sequences.

We determined new alignments of PDB sequences to PFAMs if the FATCAT p-value was $\leq$ 0.001 and the new alignment to the PFAM was longer than the original one. The new PFAM alignment is determined transitively. The procedure is as follows:

(A to B) + (B to C) = (A to C)

A to B is the HMM-exemplar alignment, B to C is the sequence alignment derived from the structure alignment of the exemplar structure to the poor PFAM domain structure, and A to C is the resulting HMM to poor PFAM domain alignment. We were able to improve the domain boundaries of 11,961 PFAM assignments. Table 3 gives the number of residues aligned by HMMER 3.0 and by FATCAT structure alignments and exemplars. The number of PDB residues aligned to PFAM HMMs is 33% higher, and the number of aligned HMM positions is 29% higher. An example of extending PFAM assignments using FATCAT structure alignment is shown in Figure 3.

A greedy algorithm for determining PFAM architectures. We developed a greedy algorithm for assigning multiple PFAMs to multi-domain protein structures in the PDB. Such an assignment is referred to as an "architecture" by PFAM itself. We used both the sequences in the PDB and the full Uniprot sequences from which these PDB sequences were derived as queries for HMMER, PSI-BLAST, and HHsearch. In some cases, the full-length sequences allow PFAM assignment when the truncated sequences used in structure determination do not (e.g. when peptides from large sequences are present in the structure). For each query, the algorithm assigns the best E-value Pfam among the sequence-HMM alignments where the sequence may be the PDB sequence, its full Uniprot sequence, or the consensus sequence of either. This region is then blocked off from further hits. If another Pfam aligns to the query and overlaps less than 10 amino acids with the already-assigned Pfams, then it will be assigned. This is performed up to an E-

value of 1.0e-5. Then the same procedure is applied to the HHsearch hits, and finally "weak hits" with poor E-values but structure alignments with p-values < 1e-3 are assigned. For proteins with repeats assigned by this procedure, any further repeats in the same Pfam are also assigned up to an E-value of 1.0.

Using the original and consensus sequence alignments to the PFAM HMMs, the HMM-HMM alignments, as well as the structure alignments, we start with the highest scoring PFAM and assign that first. If a region of the PDB sequence of the first assignment longer than 30 residues is not aligned to HMM positions, then this region is labeled "unassigned" and the assignment is deemed a "split" assignment. That is, it has a long insertion in the PDB sequence that is not part of the HMM. We assign subsequent PFAMs if they do not overlap by more than 10 residues of any existing assignment. These assignments can follow into the regions relabeled "unassigned" in the previous step. If two (or more) assignments are to the same PFAM and they cover different regions of the PFAM (allowing short overlaps), these are also considered "split" assignments. A single PFAM HMM may cover more than one sequence in a PDB entry, if the regions covered in the HMM do not overlap by more than 10 residues). We also assign repeats at weaker E-values if the sequence contains the same repeat PFAM at significant E-values (≤10e-05).

This algorithm is simple but more robust than the one PFAM itself uses. In v. 25 of PFAM, which covers PDB entries deposited through March 2011, PFAM assigns 85,446 domains to the PDB, including 3,216 repeats. Using our algorithm and the consensus and structure alignment data, we assign 107,767 domains (+26%) including 6,267 repeats (+95%) to the same PDB sequences (i.e. those available in March 2011).

Our assignments used 367 additional PFAM A models than PFAM itself or SIFTS uses, and we were able to make assignments to 99% of PDB sequences. There are still 2,327 PDB polypeptide entity sequences which are not covered in our PFAM data set, of which 1,585 sequences are less than 20 amino acids and/or contain Unknown or non-standard amino acids. A total of 96 of these 742 sequences with length ≥ 20 amino acids do not have Uniprot codes. The majority of these come from disordered regions and engineered proteins.

The PDBfam paper was published in November 2012. We have updated the database several times, including a major update to version 27 of Pfam, which was released in March 2013. This entailed running the original PDB sequences and their consensus versions through a new set of 14,831 Pfam hidden Markov models (HMMs), as well as repeating all of the structure alignments because of the changes in definitions in Pfam domains. The current assignments include 8,522 different Pfams represented by at least one structure in the PDB, of which 6,152 have an exemplar – a hit with E-value better than 1.0e-05 and at least 80% coverage of the Pfam HMM length. The number of Pfam-Pfam interactions has increased to 3,983 for homodimer interactions and 6,165 for heterodimeric interactions, when the biological assemblies (not the asymmetric units) are used to define probable interactions.

We have recently added a new feature to PDBfam which is the annotation of ligands that exist in each file. This was a service previously provided by our ProtBUD database (Protein Biological Unit Database), but that database depended on the definition of domains by SCOP, which now covers less than half of the PDB. Rather than update ProtBUD, we merged the utility of

ProtBUD into PDBfam, so that ligands for every structure with a particular Pfam are included on the browsing page for that Pfam. An example is shown in Figure 4. The utility of this kind of information is very high. For instance, if one wants to determine the structures of CBS (cystathionine beta synthase) domains that contain adenosine analogues, following the browsing links to PDBfam's CBS domain page produces a page that shows that 44 of 117 structures contain ATP, ADP, AMP, ANP or S-adenosyl methionine. A superposition of 15 structures with AMP bound (two sites per CBS-domain tandem pair) is shown in Figure 4. Further, we have added to PDBfam an option to download coordinates and Pymol scripts to open up all of the structures in the PDB that contain a given Pfam and to align these domains within Pymol. An example is shown in Figure 5.

A snapshot from PDBfam is shown in Figure 6 for the CBS domains, including the ligands bound to CBS. This is very useful when one wants to identify a structure in a large family with a particular ligand bound or for getting an idea of the kinds of ligands that bind to a particular Pfam (ATP, DNA, etc.)

Split domains and unassigned regions. The greedy algorithm identifies large insertions in PDB sequences compared to the HMMs and unassigns these regions so that later in the process other domains may be assigned to the unaligned regions. In other cases, we have two separate alignments of different portions of the HMM to different regions of the PDB sequence with significant insertions in between. We call these "split domains." Pfam itself often annotates the latter group as two copies of the Pfam domain, when in fact is only one with a large insertion. Table 4 lists the kind of split domains we observe in the PDB. They are not a large portion of the PDB but it is nevertheless important to deal with them properly.

Figures 7 and 8 present an analysis of the unassigned regions in PDB sequences. These may be N or C-terminal segments adjacent to Pfam assignments, internal segments between Pfam assignments, or PDB sequences with no Pfam assignments. Figure 7 shows that the lengths of unassigned regions next to Pfams are mostly very short and that the 85% of completely unassigned sequences are shorter than 50 amino acids. Many of these are peptides that come from intrinsically disordered regions of proteins and they are not represented in any Pfam hidden Markov model. They are not usually assigned folds by SCOP or CATH either. Figure 8 shows that the unassigned regions tend to have a higher proportion of loop regions (i.e. not helix or sheet) and disordered regions (i.e., no coordinates present in the structure these sequence regions).

Pfam-Pfam Interactions. Pfam assignments have been used previously to catalogue the physical interaction of different domain families within the PDB. With a larger and more accurate set of assignments, we investigated the number of such interactions that are now present in the PDB. The numbers of interchain and intrachain interactions of Pfams are given in Table 5. An interaction is defined if there are at least five pairs of residues with any atomic distance < 5Å, or at least 10 pairs of residues with Cβ/Cα distance < 12Å and at least one atomic distance < 5Å. The column "FCCC All BA" gives the results for the biological assemblies in the current PDB. There are 3,499 Pfams involved in homodimeric interactions between chains (or as heterodimers of two proteins containing the same Pfam domains). There are 3,958 pairs of Pfams in physical interactions across two protein chains. If we consider interchain and intrachain relationships

together, there are 3,576 Same-pfam interactions and 6,132 Diff-pair interactions involving 3,982 Pfams.

The 3DID database (3did.irbbarcelona.org) also uses Pfam to determine the prevalence of domain interactions in the PDB. We parsed the 3DID Pfam interactions directly from the text file (3did_flat_Apr_3_2011.dat) available from 3DID. This file contains Pfam interactions for 35,449 entries. 3DID contains interactions present in asymmetric units and does not utilize the biological assemblies of the PDB. Table 5 compares our results with theirs. To accomplish this, we calculated the number of Pfam interactions in the same set of 35,449 PDB entries in both the asymmetric units (FCCC(ASU)) and in the biological assemblies (FCCC(BA)). The 3DID results are given in the last column (3DID(ASU)). Because we have more Pfam assignments for these entries, we have more Pfam interactions in the asymmetric units. However, when using the biological assemblies the number of interchain Pfam interactions (both same-pfam and diff-pfam) are reduced from those of the ASU. These results underline the importance of using biological assemblies in the analysis of protein-protein interactions in the PDB.

2. Pfam clans in the PDB

The main drawback of Pfam compared to SCOP and CATH is the problematic nature of superfamily relationships in Pfam. In SCOP and CATH, these relationships comprise domains of roughly equal size and topology with few exceptions. Some "clan" relationships in Pfam have been determined structurally, but many have been established in the absence of structure. In many cases, the HMMs in a single Pfam clan vary in length by factors of three or more, indicating they must represent rather different elements of structure. As an example, the Pfam clan CL0013, "Beta-lactamase," contains seven different Pfam domain family members for which we have structures. Six out of seven of these have a domain with a Beta-lactamase fold. The Pfam models are 200-444 amino acids long. The seventh of these Pfams, and the shortest, DAPB, with length 88 is a small 8-stranded Beta-barrel and does not share the Beta-lactamase fold. The reason it is in this clan, apparently, is that the same protein (PDB 1EI5) has an N-terminal domain with a Beta-lactamase fold with Pfam domain "Beta_lactamase." In Figure 9, we show the number of Pfam clan assignments that can be verified by HMM-HMM alignments and those that can be verified by structure alignment using the program FATCAT. The sequence-based method is statistically significant (E-value<0.001) only for about 30% of same-Clan/different-Pfam relationships, while the structure-based method verifies a little over 80% of such relationships.

Because the Pfam clans are manual annotations of very distant relationships, we hypothesized that we could model protein-protein and protein-ligand interactions at very long evolutionary distances as long as we could verify the structural similarity of proteins in the same Pfam clan but different Pfam families. We used five different structural alignment programs and hierarchical clustering to examine the structural validity of Pfam clans. Of 390 clans with more than one Pfam member in the PDB, we found that 252 of them produced completely consistent structure alignments across the five programs. Two of these are shown in Figure 10. But we also found a number of issues in the remaining 138 clans that need to be dealt with before the Pfam clans can be used for structure prediction.

Many members of the same clan are of quite different length and this appears to be for two

reasons. The simplest is that some clans contain members that are structural repeats of other members of the same clan. We found 35 clans contain different numbers of the same kind of domain in their Pfam definitions. Once these are catalogued, each individual domain can be used as a template for protein-protein and protein-ligand modeling separately. Second, many Pfams contain auxiliary domains that other members of the same clan do not, even though they share a core domain that has roughly the same fold. Sometimes these domains are small N or C terminal extensions of the central core domain but in other cases they are fully-fledged domains. This occurs in 43 clans, and two of these are shown in Figure 11. In such cases, it is important to use only the shared core domains to model molecular interactions and not the auxiliary domains that are not shared. Third, several clans have what we describe as incomplete graph relationships. In these cases, one Pfam member is a two-domain protein, while other clan members are similar to one domain or the other but not both. The 1-3 relationship can therefore not be used for inference about protein interactions since the domains are evolutionarily and structurally unrelated. There are five such examples, and one of them is shown in Figure 12. Finally, there are a number of clans that contain one or more members that do not fit since they are not structurally related and share no elements of similar structure. There are 55 examples of this.

An example of how the clan relationships can be useful is shown in Figure 13, which shows several members of the HMG-box (high-mobility group) clan which bind DNA or RNA in the structures of some Pfams. A total of 101 clans have DNA or RNA in one or more of their structures. We can extend these relationships to how other members of the same clan also bind DNA in the absence of a co-crystal structure. A paper on the work on Pfam clans has been written and is currently being edited.

3. ProtCID on domain level

The original version of ProtCID – the Protein Common Interface Database published in January 2011 was performed at the chain level. That is, protein-protein interactions in all protein crystals were clustered at the level of entire chains. So multi-domain proteins are only clustered if they contain all of the same domains. For example, proteins with (SH2)_(Pkinase) architectures were not clustered with Pkinase-only proteins. We have now completed the clustering of all Pfam-Pfam interactions at the domain level. It turns out that it was a good decision to cluster at the whole-chain level first, since the domain-level ProtCID presented a number of complications. The first is that Pfam domains do not always cover the whole structure, either because alignments are too short or the Pfam domain definitions are two short. The absence of certain structural segments makes structure alignment difficult or ambiguous. The number of structure alignments that need to be performed is also very large. Finally, the display of database results and the presentation of coordinates and Pymol sessions have been revamped for the domain-level ProtCID.

There are a total of 6,962 PFAM A (v27.0) families represented in protein sequences in the PDB, of which 6,398 PFAMs have interactions with itself (same-Pfam) and other PFAMs (diff-Pfam). Table 6 summarizes the domain and chain interfaces data in ProtCID. The entries of the PDB are grouped into relations based on PFAM-PFAM interactions. Same-PFAM relations contain homodimers and some heterodimers in the case where two domains are not identical but belong to the same Pfam. In this way, there are 7070 relations with same PFAMs and 17246 relations with different PFAMs, for a total of 24316 relations from crystal structures in the PDB. There are

7,901 PFAM relations with common interfaces in more than one crystal form, comprising 65,299 distinct PDB entries and 31,731 clusters. A total of 34% of these clusters (10,908) occur between different PFAMs. There are 679 diff-PFAM clusters and 705 same-PFAM clusters presenting in at least 10 CFs with minimum sequence identity < 90%. The chain ProtCID only contains 742 clusters in the same conditions (Table 6), of which 150 clusters are heterodimeric interfaces showing the interactions between different chain PFAM architectures. Examples are shown in Figure 14.

Peptide interactions. Another addition to ProtCID is the identification of common peptide-domain interactions in the PDB. We developed a method for clustering the interactions of peptides bound to specific kinds of proteins, as defined by their Pfam domains. For every Pfam that exists in multiple structures with bound peptides, we identify the Pfam positions that are in contact (<5Å between atoms) with any residue of the peptide. We matched the residues in each peptide with those of other peptides bound to the same Pfam if they contacted at least three of the same match positions in the HMM. In order to align the peptides, we used a dynamic programming algorithm similar to a Smith-Waterman sequence alignment with an infinite gap penalty and a scoring of $\log(d_{ij}/3)$ where d_{ij} is the distance between residue i of peptide 1 and residue j of peptide 2. The alignment is local so that if the sum along a diagonal ever falls below 0 the alignment is terminated. Examples are shown in Figure 15. We then performed a search to determine whether there are protein-protein interactions (either homodimeric or heterodimeric) that are similar to known peptide/protein interactions in the PDB. These structures demonstrate the peptide-binding capacity of proteins *in situ* as segments of whole protein domains. A superposition of PDZ domains binding other proteins is shown in Figure 16.

One important use of ProtCID is in hypothesis generation of protein-protein interactions in crystals that may be biologically relevant. See below for an example. A paper on domain-level ProtCID including data on peptides and ligands and hypothesis generation is in preparation for *Nature Methods*.

BioAssemblyModeler (BAM)

Many if not most proteins function in oligomeric assemblies of one or more protein sequences. The Protein Data Bank provides coordinates for biological assemblies for each entry, at least 60% of which are dimers or larger assemblies. We have created BioAssemblyModeler (BAM), a graphical user interface to the basic steps in homology modeling of protein homooligomers and heterooligomers from the biological assemblies provided in the PDB. BAM takes as input up to six different protein sequences and begins by assigning Pfam domains to the target sequences. The program utilizes PDBfam to obtain templates that contain any or all of the domains assigned to the target sequence(s). The contents of the biological assemblies of potential templates are provided, and alignments of the target sequences to the templates are produced with a profile-profile alignment algorithm. BAM provides for visual examination and mouse-editing of the alignments supported by target and template secondary structure information and a 3D viewer of the template biological assembly. Side-chain coordinates for a model of the biological assembly are built with the program SCWRL4. A built-in protocol navigation system guides the user through all stages of homology modeling from input sequences to a target complex 3D structure. This program has been published in *PLOS* ONE and is currently available on our website, http://dunbrack.fccc.edu/BAM. A grant has been funded by NIH to develop a web version of this

methodology. A screenshot of BAM is shown in Figure 17.

Kinase autophosphorylation and regulation

Many protein kinases are activated via autophosphorylation on one or more sites. Several kinase structures have been presented in which the authors observed that the side chain of a known autophosphorylation site of one monomer in the crystal was visible in the active site of another monomer in the crystal, in a manner consistent with a potential phosphorylation event. Using a structural bioinformatics approach developed within ProtCID, we have identified several new cases of potential autophosphorylation interfaces in the PDB, some in homologous sites to those previously identified (and therefore in ProtCID clusters) and some in new sites. The current list of 14 autophosphorylation complexes (as of July 2014) are listed in Table 7. Two of the potential autophosphorylation interfaces we have identified are entirely novel observations. One of these is Y1166 of IGF1R (PDB entry 3LVP), which is temporally the second autophosphorylation site in IGF1R shown in Figure 18. This structure is asymmetric and different from the autophosphorylation structure of Y1165 of IGF1R which occurs first. The second is human CLK2 S142 (PDB 3NR9, not published), which is near the nuclear localization signal and has been found to affect CLK2 nuclear transport. Two others are similar to ones previously noted for other kinases (LCK, CSF1R).

One particularly important example of an autophosphorylation interface not previously identified is LCK in PDB entry 2PL0 (Figure 19). This is a symmetric complex of the Tyr of the activation loop of one monomer in the active site of another, and is in the same ProtCID cluster as IGF1R (PDB entry 3D94). We have generated a series of mutations in surface exposed residues at the putative LCK autophosphorylation interface. Myc-tagged wildtype or mutant LCK was transiently transfected into HEK293 cells and cell lysates were analyzed by Western blotting with either anti-Myc antibodies, to assess LCK expression levels, or anti-phospho-LCK, directed against the activation loop autophosphorylation site of LCK. The main contact between the monomers of LCK in PDB entry 2PL0 is the N-terminal region of the G-helix and the activation loop of the other monomer. We mutated N446D of the G-helix because N446 makes a hydrogen bond with a backbone carbonyl of activation loop T395. We chose T445V because this Thr forms a helix N-cap hydrogen bond to the third residue of the G-helix (Q448). Val is highly disfavored in this position. Both N446D and T445V showed strikingly reduced LCK autophosphorylation compared to wildtype (Figure 20), consistent with a role in autophosphorylation. Interestingly P447G, located in the first turn of the G-helix, markedly enhanced LCK autophosphorylation by more than 5 fold, also consistent with a role for this region in mediating autophosphorylation. The mutation P447L has been found in a leukemia cell line as an activating and transforming mutation. We speculate that the additional flexibility of the first turn of the G-helix allows for the optimization of the interaction of the G-helix with the activation loop of the other monomer.

We have analyzed the interfaces involved in autophosphorylation extensively as part of the paper on these structures. A selection of these is shown in Figure 21, with the interaction surfaces of the enzyme and substrate kinases shown in magenta on each surface. Residues in the substrate kinase about the phosphorylation position that are in contact with the enzyme kinase are shown in yellow in Table 8, which presents an alignment of the substrate sequences for the autophosphorylation complexes.

This work is in preparation for *Science Signaling*.

Hypothesis generation using ProtCID for regulation of protein kinases. One of the main uses of ProtCID at the chain and domain levels is to generate hypotheses on biologically relevant interactions in protein crystals that might not be readily apparent when the structures were determined. There are many clusters in the database that contain many crystal forms and different (but homologous proteins) for which few if any of the dimer interactions are annotated as biological. We have examined quite a few of these and over time we will work with experimental biologists to test the hypotheses generated. As an example, in Figure 22 we show a large cluster of tyrosine kinases that includes the BRAF homodimer that has been validated as an activating interaction in BRAF and has played a role in understanding the cancer-related effects of activating mutations in BRAF (e.g. V600E) and the mechanisms of BRAF inhibitors. But there are several other kinases in the same cluster, including CRAF, MLKL, CSK, ITK, and RIPK kinases. The dimer interfaces are very similar and involve some conserved residues. It is then a reasonable hypothesis that some or all of these kinases use this dimer in their biological regulation. It remains an intriguing hypothesis until there is experimental data supporting it, which we are trying to find experimental collaborators to test. In other case, a particular human serine/threonine kinase has a common dimer found in 15 different crystal forms, which is a very large number. My colleague Erica Golemis is now testing whether the kinase uses this asymmetric dimer for its regulation. So far, the experiments indicate that mutations in the interface significantly increase autophosphorylation of this kinase. Crosslinking experiments will be performed to determine if the dimer forms or whether other proteins interact at the site on one side or the other of the interface.

Table 1. Summary of PFAM A assignments from the PDB and the consensus sequences with E-values ≤ 0.001.

	PDB-orig	PDB-consensus
#PFAM A	6,090	6,461 (↑6.1%)
#Entries	69,816	70,150 (↑0.5%)
#Unique sequences	45,805	46,516 (↑1.5%)
#Domains	60,547	63,869 (↑5.5%)
#Residues	8,740,432	9,047,704(↑3.5%)
#HMM positions	8,696,426	9,017,705 (↑3.7%)

Table 2. Domain coverage of the PDB in various sources

Data Source	#entries[1]	#unique[2]	#residues[3]	#PFAM	#Repeats
PDB	79,556 (100%)	51,155 (100%)	12,781,567 (100%)	-	
Pfam-FCCC	78,649 (99%)	50,833 (99%)	10,499,637 (82%)	6,272	3,598
Pfam (v.26)	34,188 (45%)	23,621 (46%)	4,925,947 (39%)	4,874	1,173
Pfam (SIFTS)	73,009 (92%)	45,202 (88%)	9,150,584 (72%)	5,612	2,604
CATH (v.3.5)	51,334 (65%)	30,557 (60%)	6,906,806 (54%)	-	2,798
SCOP (1.75)	49,215 (62%)	29,129 (57%)	6,912,489 (54%)	-	

1. The number of PDB structures with polypeptide sequences.
2. Number of unique PDB sequences with length > 50
3. The number of residues in the covered regions of unique sequences.

Table 3. The number of residues before and after structure alignments for PDBs with weak PFAM assignments.

	#aligned Residues	#Envelope Residues	#HMM Positions
Before	1,227,544	1,428,456	1,246,328
After	1,636,550	1,636,550	1,605,329
Increased percentage ↑	33.3%	14.6%	28.8%

Table 4. The split domains in Pfam.

Format	#Domains	#Pfam	Pfam Example (HMM)
X[s1-e1]_Y_X[s2-e2]	1,018	X=93,Y=112	(ADK[1-122])_(ADK_lid)_(ADK[123-150])
X[s1-e1]_(#)_X[s2-e2]	2,366	467	(Hpt[1-69])_(35)_(Hpt[70-88])
X[s2-e2]_X[s1-e1]	352	70	(CIMR[46-145])_(CIMR[5-48])
Multi-chain domains	939	99	((2)Trypsin[1-132])(3)(Trypsin[135-220])

Table 5. Number of Pfams and Pfam pairs involved in interactions in the PDB.

		FCCC All BA	FCCC BA	FCCC ASU	3DID ASU
Inter-Chain	Same-Pfam	3,499	2,839	3,338	3,087
	Diff-Pfam Pairs	3,958	3,267	3,455	2,306
Intra-Chain	Same-Pfam	383	293	293	249
	Diff-Pfam Pairs	3,240	2,236	2,236	1,271
Same-Pfams (Inter and/or Intra)		3,576	2,909	3,382	3,127
Diff-Pfam Pairs (Inter and/or Intra)		6,132	4,720	4,796	3,049
#Pfam in diff interactions		3,982	3,336	3,338	2,658
#Pfams in same or diff interactions		5,263	4,445	4,713	4,265
#Entries		47,458	35,449	35,449	35,449

BA=Biological assembly; ASU = asymmetric unit. 47,458 entries in current PDB ("All BA"). 35,449 PDB entries listed in 3did_flat_Apr_3_2011.dat after removing obsolete entries.

Table 6. Summary of Data in domain and chain ProtCID

	Domain			Chain		
	Same-PFAM	**Diff-PFAM**	**Total (domain)**	**Single Chain Architecture**	**Pairs of Chain Architectures**	**Total (chain)**
# relations/groups	7070	17246	24316	9955	10714	20669
M≥2	3135	4766	7901	3307	1394	4701
M≥2,seqid<90	1840	2062	3902	2100	784	2884
M≥5,seqid<90	926	988	1914	902	311	1213
M≥10,seqid<90	431	466	897	395	120	515
M≥20,seqid<90	157	178	335	144	41	185
# entries	85501	35859	87113	95443	10512	95443
M≥2	57997	26547	65299	60827	7126	63206
M≥2,seqid<90	42697	21057	51506	48887	5904	51188
M≥5,seqid<90	32083	17694	40535	34028	4491	36366
M≥10,seqid<90	22401	13891	29806	23339	3153	25234
M≥20,seqid<90	13788	9304	18861	13567	2057	15250
# clusters						
M≥2	20823	10908	31731	15919	2227	18146
M≥2,seqid<90	8589	4062	12651	9235	1159	10394
M≥5,seqid<90	1999	1542	3541	1680	401	2081
M≥10,seqid<90	705	679	1384	592	150	742
M≥20,seqid<90	240	242	482	207	45	252

Table 7. Autophosphorylation interfaces identified in the PDB by ProtCID

Uniprot ID	PDB	Area	Enz/Substr	Dist	Asp	Y/ST	Location
KIT_HUMAN	1pkg	940	B:2_655; A	3.30	D792	pY568	juxtamem
CSF1R_HUMAN	3lcd	843	A:4_555; A	2.62	D778	Y561	juxtamem.
FGFR1_HUMAN	3gqi	824	A; A:4_456	3.00	D623	Y/F583	kinase insert
FGFR3_HUMAN	4k33	843	A; A:3_645	3.35	D577	Y617	kinase insert
IGF1R_HUMAN	3d94	997	A:4_555;A	2.51	D1135	Y1165	act. loop
LCK_HUMAN	2pl0	878	A:2_655; A	2.91	D364	Y394	act. loop
IGF1R_HUMAN	3lvp	848	C; B	3.14	D1135	Y1166	act. loop
EGFR_HUMAN	4i21	792	A;B	2.47	D837	Y1016	C-term
FGFR2_HUMAN	3cly	897	A:3_655; A	2.82	D626	Y769	C-term
CLK2_HUMAN	3nr9	1501	B:5_555; C	5.36	D290	S142	near NLS
KCC2D_CAEEL	3kk8	1310	A:5_555; A	4.19	N134	pT284	reg. seg.
KCC2D_CAEEL	3kk9	701	A:1_655; A	4.17	N134	pT284	reg. seg.
KCC2D_HUMAN	2wel	668	A:3_644; A	2.71	D136	T287	reg. seg.
PAK1_HUMAN	3q4z	875	A; B	2.80	N389	T423	act. loop

Enzyme/substrate are the chainids and symmetry operators that identify the relevant proteins. Autophosphorylation interfaces not identified by authors of structures are marked in red. All of these phosphorylation sites are annotated as autophosphorylation sites in the Uniprot entries. Dist. is the distance between the OH group of S/T/Y and the active site Asp OD1 or OD2. Interfaces are grouped by similarity of autophosphorylation site and kinase/kinase interface.

Table 8. **Contacts between residues in the substrate kinase and the enzyme kinase domain.** The substrate kinase sequence around the phosphorylation site is shown for each kinase. Residues in lower case are disordered. An asterisk represents the N or C terminus of the sequence. Residues in yellow are in contact with the kinase.

Uniprot	PDB	9	8	7	6	5	4	3	2	1	0	1	2	3	4	5	6	7	8	9
KIT_HUMAN	1pkg			*	e	i	n	g	N	N	X	V	X	I	D	P	T	Q	L	P
				*	e	i	n	g	n	N	X	V	X	I	D	P	T	Q	L	P
CSF1R_HUMAN	**3lcd**			*****	**s**	**y**	**e**	**G**	**N**	**S**	**Y**	**T**	**F**	**I**	**D**	**P**	**T**	**Q**	**L**	**P**
IGF1R_HUMAN	3d94	M	T	R	D	I	Y	E	T	D	Y	Y	R	K	G	G	K	G	L	L
LCK_HUMAN	**2pl0**	**L**	**A**	**R**	**L**	**I**	**E**	**D**	**N**	**E**	**Y**	**T**	**A**	**R**	**E**	**G**	**A**	**K**	**F**	**P**
IGF1R_HUMAN	**3lvp**	**T**	**R**	**D**	**I**	**Y**	**E**	**T**	**D**	**Y**	**Y**	**R**	**K**	**G**	**G**	**K**	**G**	**L**	**L**	**P**
FGFR1_HUMAN	3gqi	Q	A	R	R	P	P	G	L	E	F	S	F	N	P	S	H	N	P	E
FGFR2_HUMAN	3cly	I	L	T	L	T	T	N	E	E	Y	L	D	L	s	q	p			
EGFR_HUMAN	4i21	S	P	T	V	V	D	A	D	E	Y	L	I	P	Q	q	q			
		S	P	T	V	V	D	A	D	E	Y	L	I	P	Q	q	q			
PAK1_HUMAN	3q4z				e	Q	S	K	R	S	T	M	V	G	T	P	Y	W	M	A
KCC2D_CAEEL	3kk8	R	V	A	A	I	H	R	Q	D	T	V	D	C	L	*				
KCC2D_CAEEL	3kk9	R	V	A	A	I	H	R	Q	D	Y	V	D	C	l	*				
KCC2D_HUMAN	2wel	Q	R	S	M	M	H	R	Q	E	T	V	D	C	L	K	K	F	N	A
CLK2_HUMAN	**3nr9**			*****	**S**	**S**	**R**	**R**	**A**	**K**	**S**	**V**	**E**	**D**	**D**	**A**	**E**	**G**	**H**	**L**

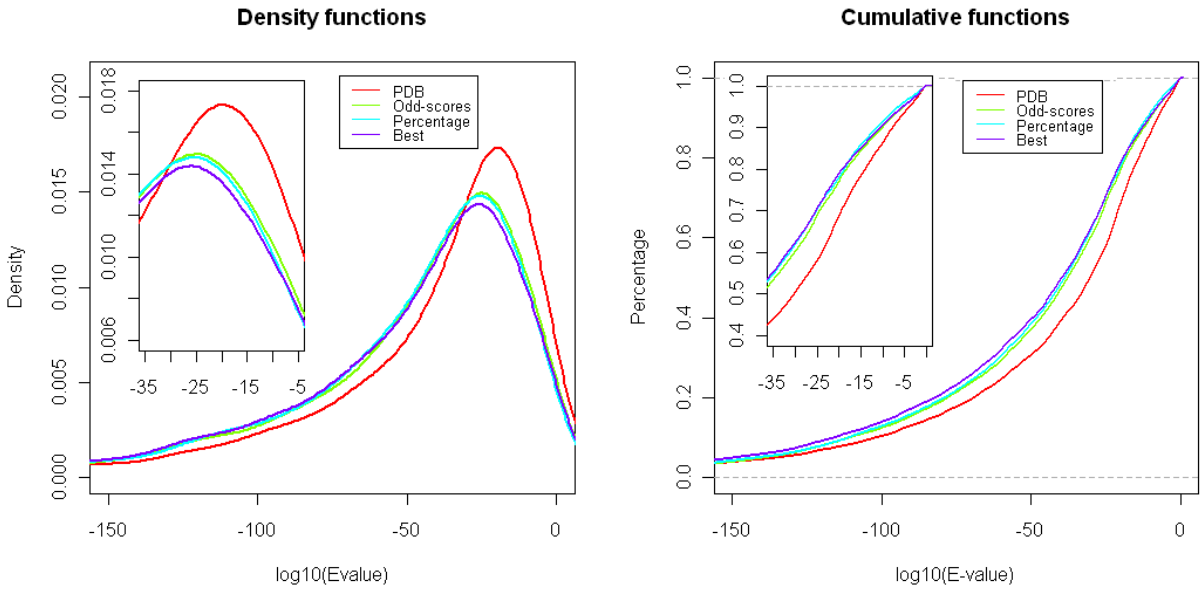

Figure 1. Comparison between PDB sequences and consensus sequences. Left: density functions. Right: cumulative density functions.

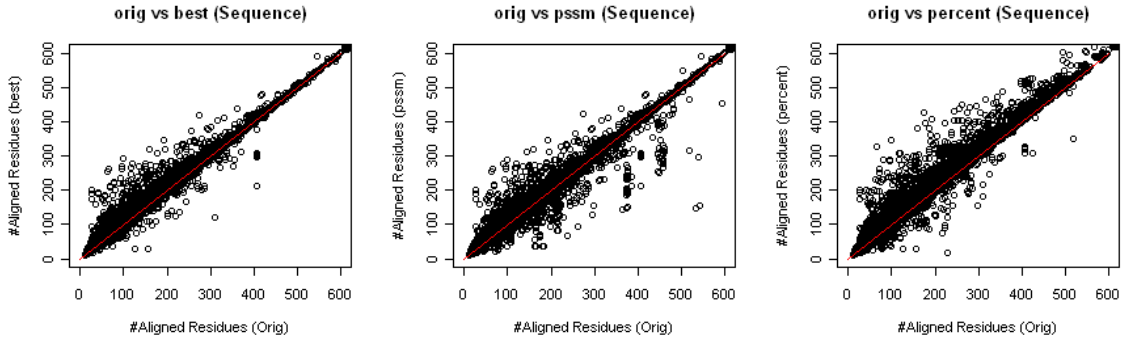

Figure 2. Scatter plots of the numbers of aligned residues of the original PDB sequences versus the numbers of aligned residues of the consensus sequences and PDB-best.

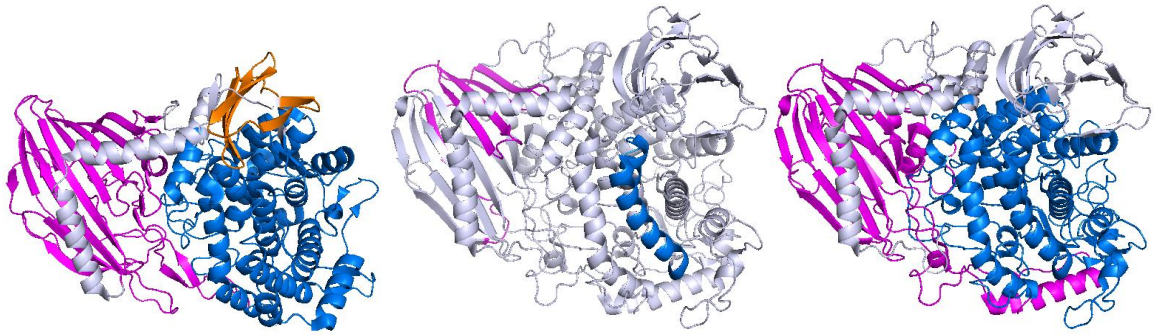

Figure 3. Structure alignments expand the PFAM assignments. Left: PDB entry 1H54 has three PFAM assignments in magenta, blue and orange. Center: PDB entry 2EAB has two PFAM assignments with weak E-values and short alignments, in magenta and blue. These assignments are in the same clans as those in 1H54 of the same colors. Right, after structure alignment of 2EAB and 1H54, much longer PFAM assignments of these two domains can be made.

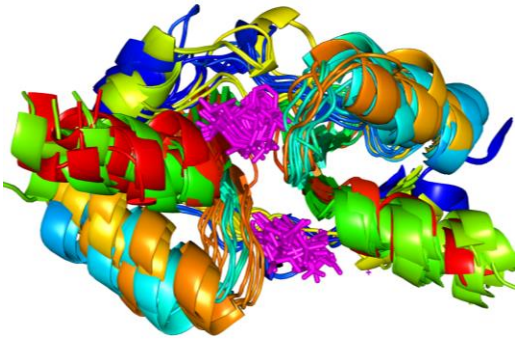

Figure 4. CBS domains bound with AMP

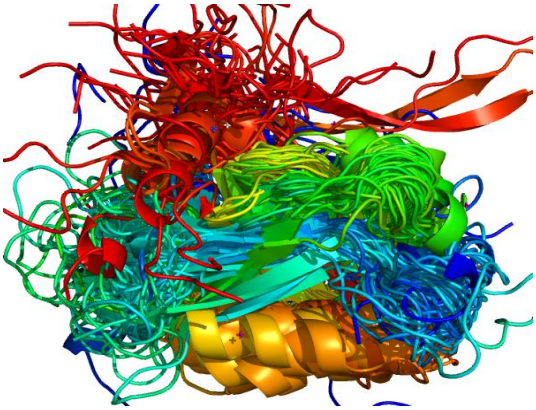

Figure 5. PDBfam alignment of PDZ domains

PDBID	EntityID	Chain Arch	E-Value	SeqBeg	SeqEnd	HmmBeg	HmmEnd	Entry Arch	Unp Codes	PDBBUs	PISABUs	Ligands
4hg0	1	(CBS)_(CBS)_(CorC_HlyC)	5.4E-014	69	128	1	55	(CBS)_(CBS)_(CorC_HlyC)	(CORC_ECOLI)	A2	A2	(AMP)
4hg0	1	(CBS)_(CBS)_(CorC_HlyC)	1.2E-013	135	190	2	56	(CBS)_(CBS)_(CorC_HlyC)	(CORC_ECOLI)	A2	A2	(AMP)
4gqw	1	(CBS)_(CBS)	6.5E-026	87	143	1	56	(CBS)_(CBS)	(CBSX1_ARATH)	A2	A2	
4gqw	1	(CBS)_(CBS)	2.0E-024	7	65	1	53	(CBS)_(CBS)	(CBSX1_ARATH)	A2	A2	
4gqv	1	(CBS)_(CBS)	7.8E-026	101	157	1	56	(CBS)_(CBS)	(CBSX1_ARATH)	A2	A2	
4gqv	1	(CBS)_(CBS)	1.0E-023	7	65	1	54	(CBS)_(CBS)	(CBSX1_ARATH)	A2	A2	
4fxs	1	(IMPDH[2-85])_(CBS)_(CBS)_(IMPDH[86-352])	2.5E-015	91	145	4	52	(IMPDH[2-85])_(CBS)_(CBS)_(IMPDH[86-352])	(Q9KTW3_VIBCH)	A4	A8	(IMP)(K)(MOA)
4fxs	1	(IMPDH[2-85])_(CBS)_(CBS)_(IMPDH[86-352])	3.5E-015	151	209	1	54	(IMPDH[2-85])_(CBS)_(CBS)_(IMPDH[86-352])	(Q9KTW3_VIBCH)	A4	A8	(IMP)(K)(MOA)
4fry	1	(CBS)_(CBS)	1.9E-021	80	135	1	55	(CBS)_(CBS)	(B1YXI0_BURA4)	A2	A2	(AMP)(NAD)
4fry	1	(CBS)_(CBS)	5.6E-020	6	70	2	56	(CBS)_(CBS)	(B1YXI0_BURA4)	A2	A2	(AMP)(NAD)
4esy	1	(CBS)_(CBS)	1.2E-025	20	75	1	54	(CBS)_(CBS) (CBS)_(CBS)	(D1CAK9_SPHTD) (D1CAK9_SPHTD)	AB	AB	(CL)(EDO)

Figure 6. PDBfam table for the Pfam "CBS," including the ligands present in each structure. These proteins bind adenosine analogues, so the table assists in identifying structures with these ligands (AMP, ATP, etc.)

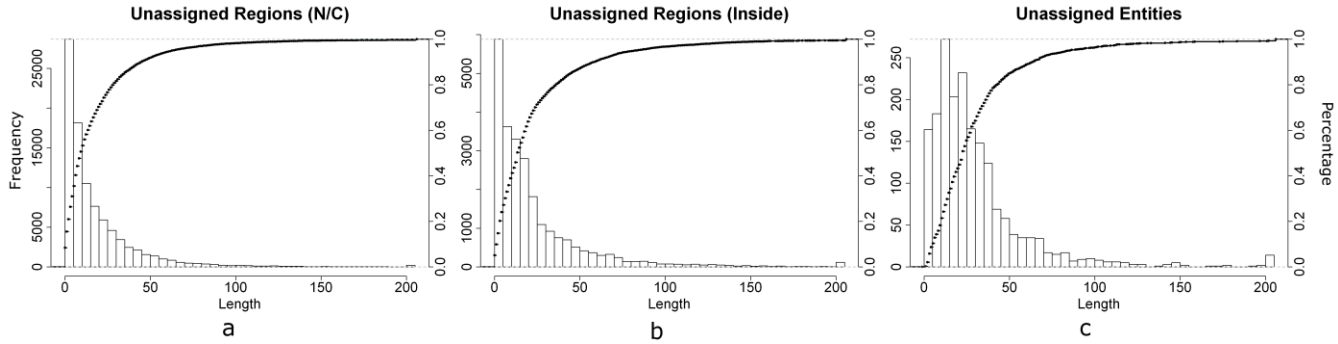

Figure 7. Histograms of the lengths of the unassigned regions/entities. (a): unassigned regions at N/C terminals. (b): unassigned regions between two Pfam assignments. (c): entities with no Pfam assigned. The curves represent the cumulative percentage as a function of increasing length (right axis).

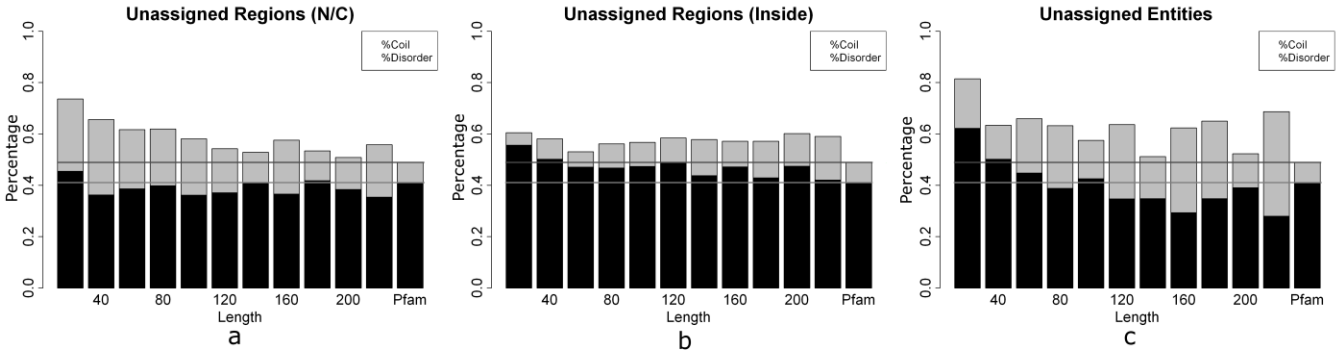

Figure 8. Secondary structure of the unassigned regions/sequences and Pfam assignments. Disorder means the residues do not have coordinates in the PDB file, while coil includes all residues that are not α-helix and or β-strand. The y-axis is the number of coil + disordered residues divided by the total unassigned residues at that length range. A bar is 20 amino acids long, i.e. 1-20, 21-40, 41-60, etc. The last bar in each figure shows the values for Pfam assignments. The next to last column in each figure shows the values for all regions with length > 200. The light gray horizontal line is the rate of coil in the Pfam assigned regions, and the dark gray horizontal line is the rate of coil+disorder in the Pfam assignments.

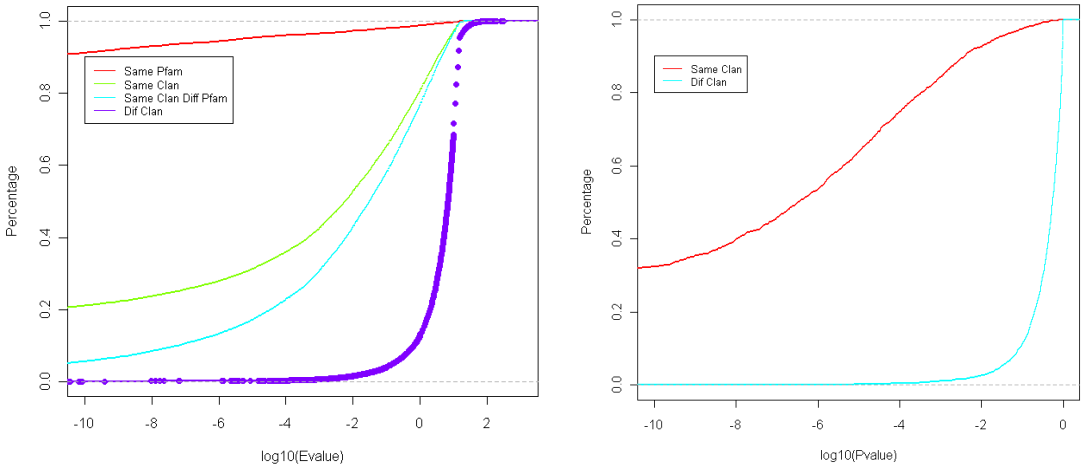

Figure 9. Cumulative distributions of HHSearch E-values and FATCAT structure alignment p-values for same-clan and different-clan relationships in Pfam.

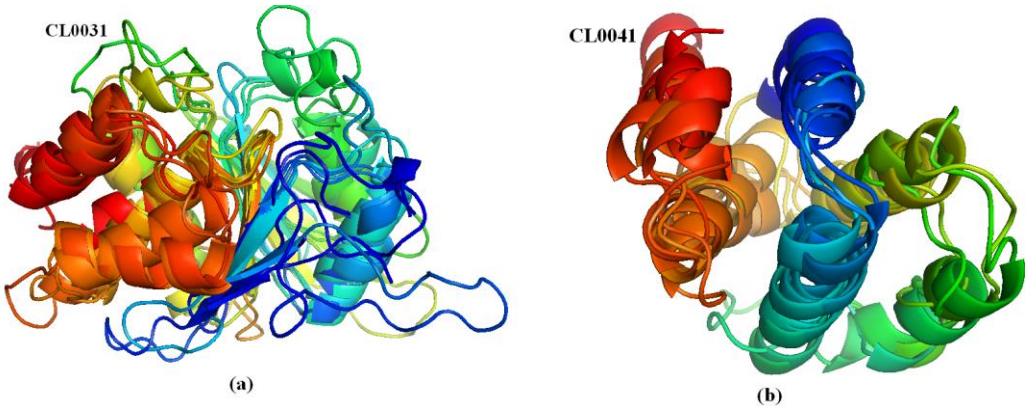

Figure 10. Multiple structure alignments for two PFAM clans which contain structurally related PFAM members. (a) multiple structure alignment of clan CL0031; (b) multiple structure alignment of clan CL0041.

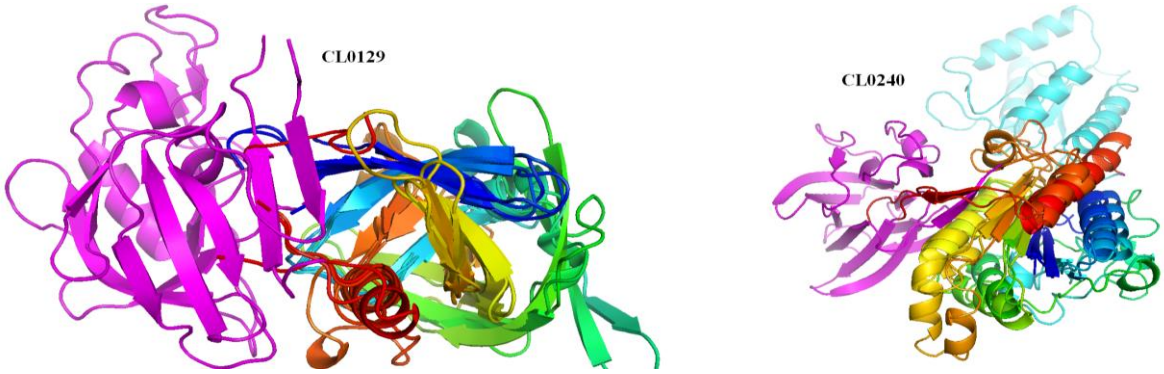

Figure 11. Pfams with extra domains (magenta) not present in most members of the same clan (rainbow).

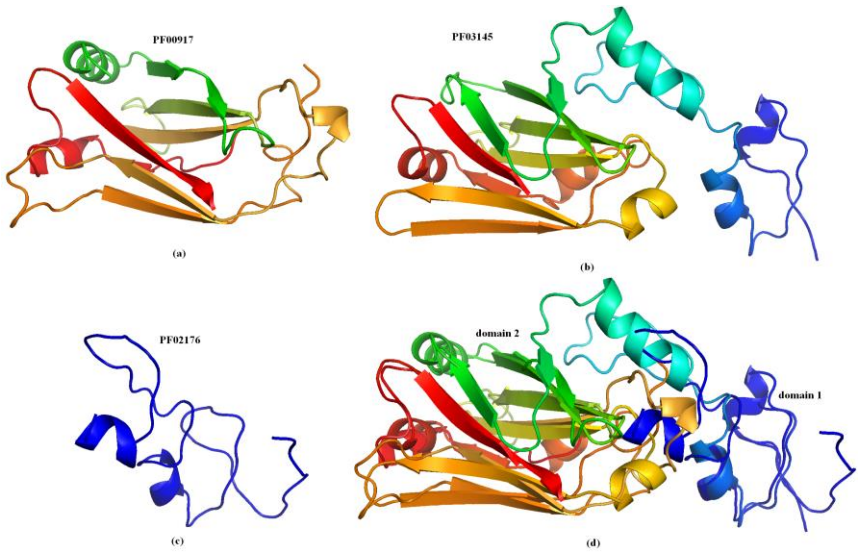

Figure 12. Example of incomplete graph relationship. The multiple structure alignment of three PFAM members in clan CL0389. (a) structure of PFAM PF00917; (b) the structure of PFAM PF03145; (c) the structure of PF02176; (d) the multiple structure alignment of these three PFAMs in clan CL0389.

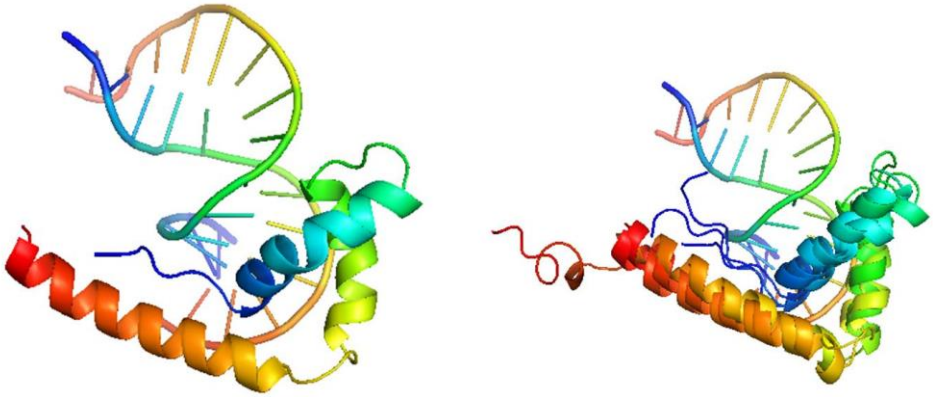

Figure 13. Using verified Pfam clan relationships to model protein/DNA interactions. The structure on the left is of one member of clan 114 bound to DNA Two other Pfam members of the same clan do not have structures that contain DNA but can be modeled bound to DNA by superposition (right figure).

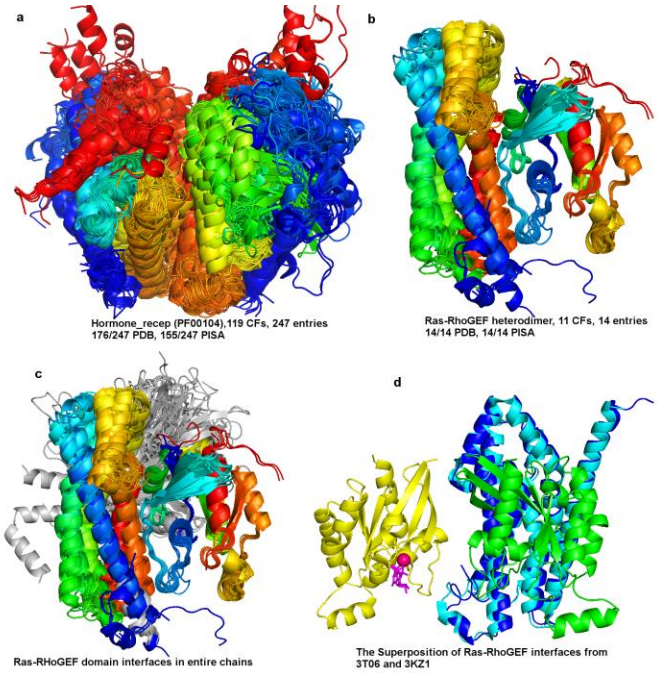

Figure 14. Domain Interface Clusters in ProtCID. (a). Same-PFAM dimers of Hormone_recep (PF00104). The biggest cluster of domain interfaces present in 119 CFs and 247 PDB entries (only one interface per CF is shown), including 217 homodimers and 30 heterodimers from 11 different species and 33 distinct UniProt sequences (20 human sequences). (b). Diff-PFAM Domain interfaces in ProtCID. Heterodimers of PFAM domains Ras (small GTPase), and RhoGEF (DH domain), present in 11 CFs and 14 PDB entries. Ras is in the left of the interface and RhoGEF is in the right. (c). The Ras and RhoGEF heterodimers in the chain-based interfaces. PFAM domains are colored from blue to red from N- to C-terminus respectively; the other parts of chains are colored in grey. Domains are aligned with PyMOL to one representative interface with best resolution. (d). The superposition of Ras-RHoGEF interfaces from 3T06 and 3KZ1. RhoA Ras and RhoGEF domains in 3T06 are colored green and blue respectively. The domains in 3KZ1 are colored yellow and cyan respectively. GSP is colored magenta and shown as sticks. Magnesium ion is colored pink and shown as a sphere. RhoGEF domains are superposed.

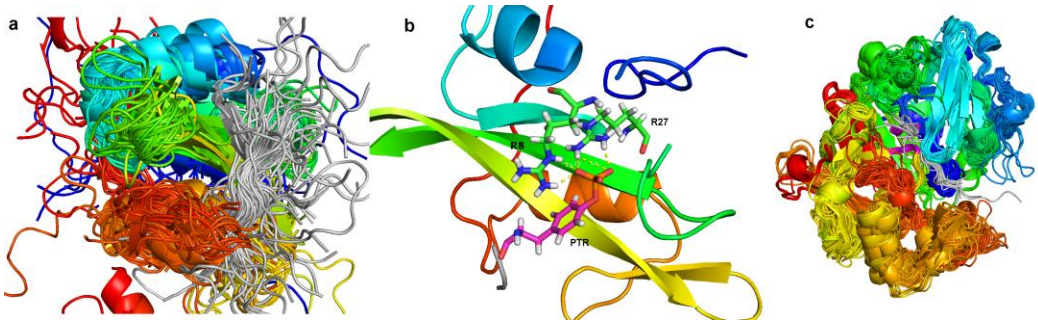

Figure 15. PFAM-peptide interfaces. (a). The cluster of SH2 and peptide interfaces contains 99 peptide dimers. (b). The Arg8 and Arg27 residues form hydrogen bonds with phosphotrysine in 1LKK. The numbers 8 and 27 are SH2 HMM positions. Arginine residues and PTR are shown as sticks. (c). The cluster of Asp (Aspartate protease) and peptide interfaces, consists of 58 interface complexes. The Asp-Thr-Gly motifs are shown magenta. Domains are colored from blue to red from N- to C- terminus, and aligned by the positions of PFAM HMMs. Peptide chains are colored gray.

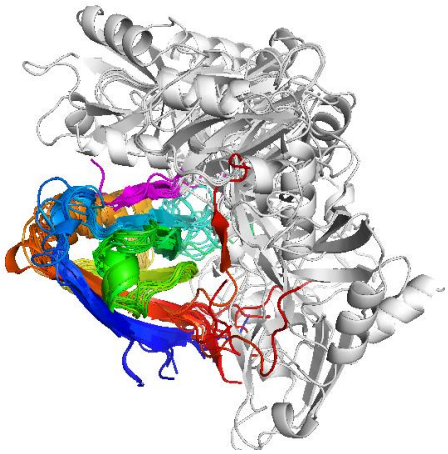

Figure 16. Proteins (in white) with their C-terminal segments (magenta) bound to PDZ domains (rainbow) in a manner resembling known PDZ/peptide interactions.

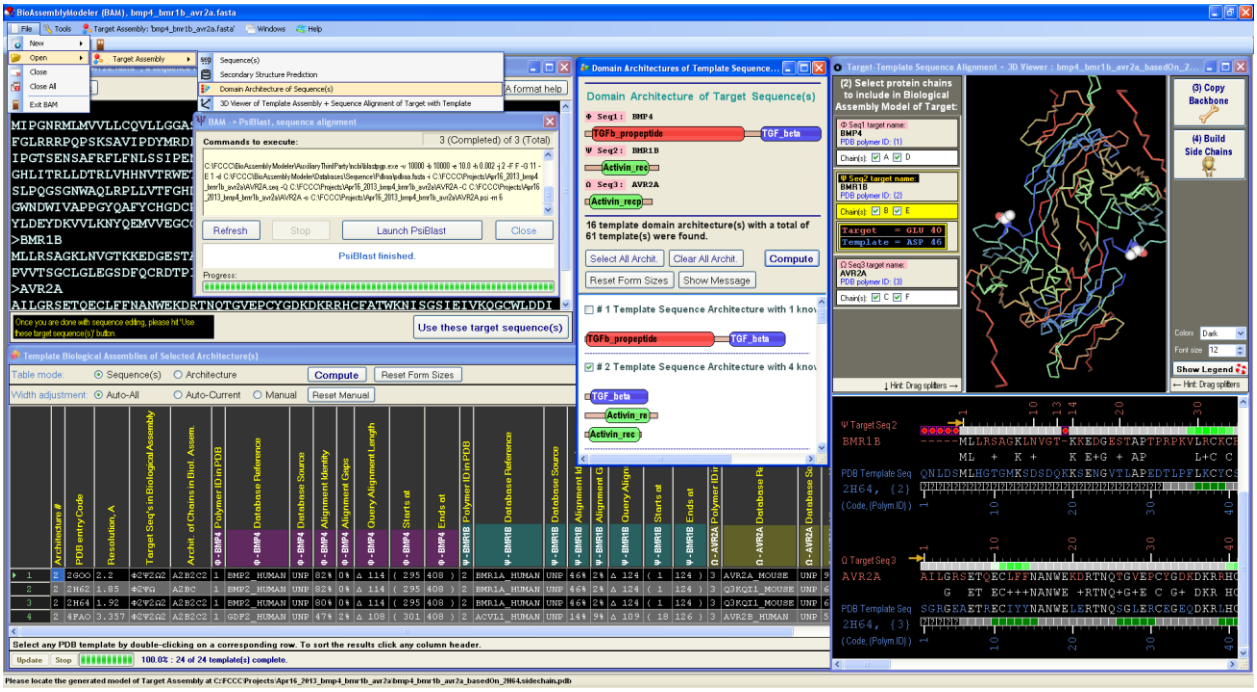

Figure 17. Screenshot of BioAssemblyModeler (BAM). The target is a heterohexameric complex of three different proteins, human BMP2 and its type I and type II receptors, BMR1B, and AVR2A.

A. **B.** **C.**

Figure 18. Autophosphorylation complex of IGF1R Tyr 1166.
A. Asymmetric unit monomer chain C (rainbow) is the enzyme kinase and monomer chain B (magenta) is the substrate. Y1165 and Y1166 of chain C is shown in orange stick figure and demonstrate that the enzyme kinase is in active conformation. Y1165 of the substrate is shown in magenta sticks.
B. Asymmetric unit monomers chains D (rainbow) and A (magenta) form the same interface as the C-B dimer in panel A, but the activation loop region around Y1166 is chain A is disordered and Y1166 is not present in the coordinates.
C. Close up of IGF1R structures of the enzymatic kinases from PDB 3D94 and 3LVP aligned with their substrate kinase loops shown in magenta (3D94) and blue (3LVP). The Tyr residues are (1165 and 1166 respectively) are shown in sticks opposite the active site Asp side chains.

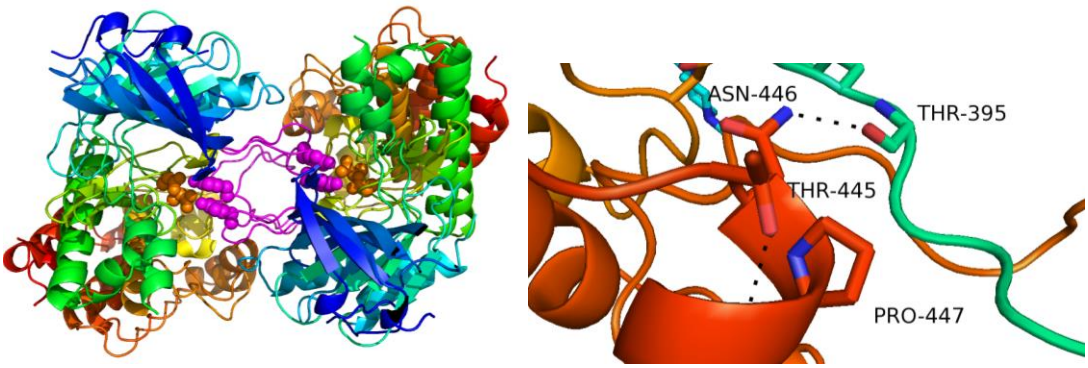

Figure 19. Kinase autophosphorylation. Left, ProtCID cluster of IGF1R (3D94:Y1165) and LCK (2PL0) dimers involving trans-autophosphorylation of the activation loop. Tyr site and active site Asp residues marked in spheres. Right, close up of 2PL0 G-helix interaction (orange) with the activation loop of other monomer

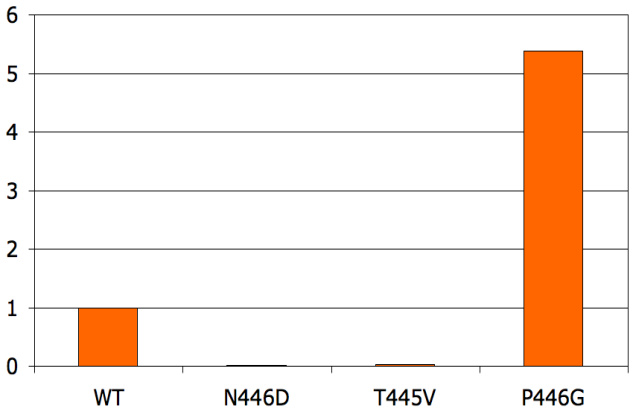

Figure 20. LCK autophosphorylation by anti-phospho-LCK binding, relative to WT.

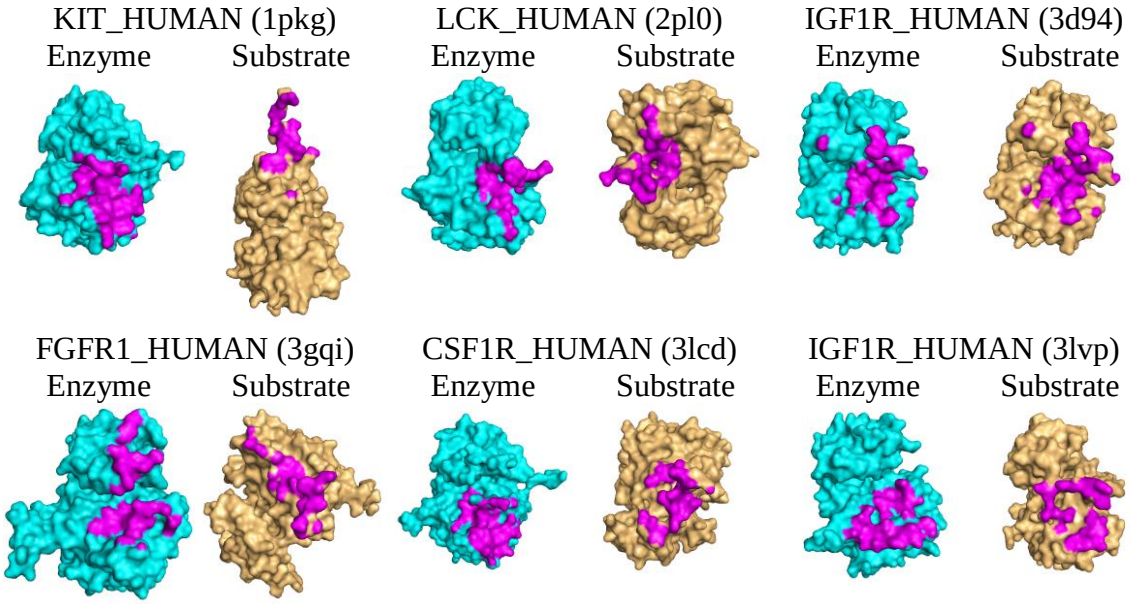

Figure 21. Interaction surfaces of autophosphorylation complexes. The residues within 5 Å of the other protein are shown in magenta on the protein surfaces. The enzyme kinases are oriented similarly across the kinases with the N-terminal domain on top and the C-terminal domain on the bottom, and the active site facing the viewer.

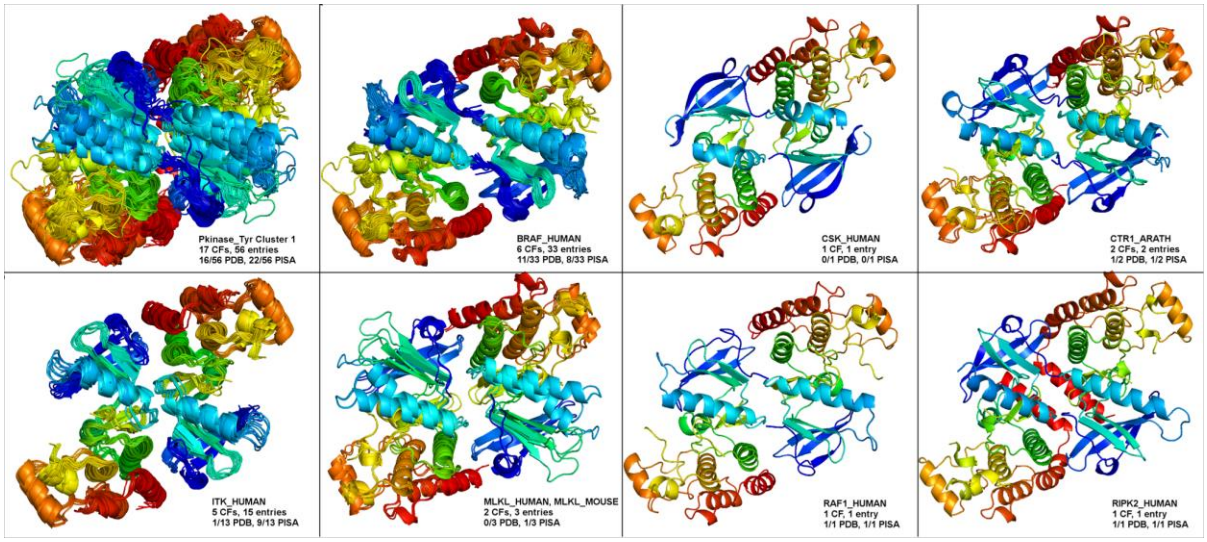

Figure 22. Tyrosine kinase activating homodimers in Pkinase_Tyr domain cluster 1 with 17 crystal forms (CFs) and 56 PDB entries. The proteins include human BRAF, CSK, CTR1, ITK, MLKL, RAF1 (CRAF) and RIPK2, as well as mouse MLKL. The domains are colored blue to red from N to C terminus. The annotation for each image includes the protein name, the number of CFs, the number of PDB entries, and the number of PDB and PISA biological assemblies that contain the observed dimer (e.g. 16 of 56 dimers in the entire cluster are annotated as biological in the PDB).

Scientific presentations using the results of work funded under this project:

1. Jul. 15, 2011 *Identifying biologically relevant interactions in protein crystals.* Intelligent Systems and Molecular Biology Conference 3DSIG Meeting, Vienna, Austria
2. Aug. 31, 2011 *Common interfaces in multiple crystal forms of proteins and protein complexes associated with biological function.* 242nd American Chemical Society Annual Meeting, Denver CO
3. Oct. 28, 2011. *Statistical analysis of protein structures: backbone conformations, side-chain rotamers and protein-protein interfaces.* The 2nd Biocomputation Forum, Asia University, Taichung, Taiwan
4. Nov. 30, 2011. *Statistical analysis of protein structures for structure prediction and protein design.* Raymond & Beverly Sackler Institute for Biological, Physical and Engineering Sciences, Yale University, New Haven CT
5. Feb. 28 2012 *Structural bioinformatics of antibodies and protein-protein interactions*. Friday Research Discussion, University of Pennsylvania, Department of Biochemistry and Biophysics, Philadelphia PA
6. Sep. 21, 2012 *Protein structure in biology and medicine.* First Annual Temple Biomedical Research Day, Senior Excellence Award Lecture. Temple University School of Medicine, Philadelphia PA
7. Oct. 4, 2012 *Structural bioinformatics of proteins and protein complexes*. Keynote lecture, Computational Structural Bioinformatics Workshop, IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2012). Philadelphia PA
8. Nov. 8, 2012 *Regulation of protein kinases through protein-protein interactions*. Conference on Modeling of Protein Interactions (MPI), Lawrence, KS
9. Dec. 3, 2012 *Structural bioinformatics and protein structure prediction*. Department of Biochemistry, Temple University School of Medicine, Philadelphia, PA
10. Dec. 7, 2012 *Structural bioinformatics and protein structure* prediction. Dipartimento di Scienze Biomolecolari, Università degli Studi di Urbino "Carlo Bo,", Urbino, Italy
11. Dec. 12, 2012 *Structural bioinformatics for protein structure prediction*. 10th Meeting on the Critical Assessment of Protein Structure Prediction (CASP10), Gaeta, Italy
12. Jan. 18, 2013 *Structural bioinformatics for protein structure prediction*. Department of Biological Sciences and Bioengineering, Indian Institute of Science, Kanpur, India
13. Mar. 13, 2013 *Structural bioinformatics for protein structure prediction*. Mid-Atlantic Computational Chemistry meeting, Bryn Mawr, PA
14. Apr. 17, 2013 *Structural bioinformatics and protein structure prediction*. Department of Biochemistry and Molecular Biology, University of Tennessee, Knoxville, TN
15. Apr. 19. 2013 *Modeling biological assemblies*. Department of Biochemistry and Molecular

Biophysics, University of Pennsylvania, Philadelphia, PA

16. Jul. 17, 2013 *Prediction of phenotypes of MRN complex mutations*. Meeting on Critical Assessment of Genome Interpretation (CAGI), Berlin, Germany
17. Jul 17, 2013 *Assessment of predictions of mutations in BRCA1 and BRCA2*. Meeting on Critical Assessment of Genome Interpretation (CAGI), Berlin, Germany
18. Jul. 19, 2013 *Autophosphorylation complexes in crystals of protein kinases*. 3DSIG: Structural Bioinformatics and Computational Biophysics Meeting, Berlin, Germany
19. Mar. 13, 2014 *Structural bioinformatics and protein structure prediction*. Department of Chemistry, Temple University, Philadelphia, PA
20. May 19, 2014 *Structural bioinformatics and protein structure prediction*. Ort Braude College, Karmi'el, Israel
21. May 21, 2014 *Structural bioinformatics for protein structure prediction and design*. Edmund J. Safra Center for Bioinformatics Distinguished Speaker Series, Tel Aviv University, Tel Aviv, Israel
22. May 27, 2014 *Structural bioinformatics for protein structure prediction and design*. Department of Structural Biology, Weizmann Institute, Rehovot, Israel
23. Jun. 1, 2014 *Structural bioinformatics of protein-protein interactions*. Faculty of Agriculture, the Hebrew University of Jersusalem, Rehovot, Israel
24. Aug. 5, 2014 *The structures of biological assemblies and predicting the effects of missense mutations*. Gordon Research Conference on Single Nucleotide Polymorphisms and Disease, Stonehill College, Easton, MA

18. Extent of Clinical Activities Initiated and Completed. Items 18(A) and 18(B) should be completed for all research projects. If the project was restricted to secondary analysis of clinical data or data analysis of clinical research, then responses to 18(A) and 18(B) should be "No."

18(A) Did you initiate a study that involved the testing of treatment, prevention or diagnostic procedures on human subjects?

_____Yes
__X___No

18(B) Did you complete a study that involved the testing of treatment, prevention or diagnostic procedures on human subjects?

_____Yes
___X__No

If "Yes" to either 18(A) or 18(B), items 18(C) – (F) must also be completed. (Do NOT complete 18(C-F) if 18(A) and 18(B) are both "No.")

18(C) How many hospital and health care professionals were involved in the research project?

_____Number of hospital and health care professionals involved in the research project

18(D) How many subjects were included in the study compared to targeted goals?

_____Number of subjects originally targeted to be included in the study
_____Number of subjects enrolled in the study

Note: Studies that fall dramatically short on recruitment are encouraged to provide the details of their recruitment efforts in Item 17, Progress in Achieving Research Goals, Objectives and Aims. For example, the number of eligible subjects approached, the number that refused to participate and the reasons for refusal. Without this information it is difficult to discern whether eligibility criteria were too restrictive or the study simply did not appeal to subjects.

18(E) How many subjects were enrolled in the study by gender, ethnicity and race?

Gender:
_____Males
_____Females
_____Unknown

Ethnicity:
_____Latinos or Hispanics
_____Not Latinos or Hispanics
_____Unknown

Race:
_____American Indian or Alaska Native
_____Asian
_____Blacks or African American
_____Native Hawaiian or Other Pacific Islander
_____White
_____Other, specify:_____
_____Unknown

18(F) Where was the research study conducted? (List the county where the research study was conducted. If the treatment, prevention and diagnostic tests were offered in more than one county, list all of the counties where the research study was conducted.)

19. Human Embryonic Stem Cell Research. Item 19(A) should be completed for all research projects. If the research project involved human embryonic stem cells, items 19(B) and 19(C) must also be completed.

19(A) Did this project involve, in any capacity, human embryonic stem cells?

_____Yes

___X_ No

19(B) Were these stem cell lines NIH-approved lines that were derived outside of Pennsylvania?

_____Yes

_____ No

19(C) Please describe how this project involved human embryonic stem cells:

20. Articles Submitted to Peer-Reviewed Publications.

20(A) Identify all publications that resulted from the research performed during the funding period and that have been submitted to peer-reviewed publications. Do not list journal abstracts or presentations at professional meetings; abstract and meeting presentations should be listed at the end of item 17. **Include only those publications that acknowledge the Pennsylvania Department of Health as a funding source** (as required in the grant agreement). List the title of the journal article, the authors, the name of the peer-reviewed publication, the month and year when it was submitted, and the status of publication (submitted for publication, accepted for publication or published.). Submit an electronic copy of each publication or paper submitted for publication, listed in the table, in a PDF version 5.0.5 (or greater) format, 1,200 dpi. Filenames for each publication should include the number of the research project, the last name of the PI, and an abbreviated title of the publication. For example, if you submit two publications for Smith (PI for Project 01), one publication for Zhang (PI for Project 03), and one publication for Bates (PI for Project 04), the filenames would be:

Project 01 – Smith – Three cases of isolated
Project 01 – Smith – Investigation of NEB1 deletions
Project 03 – Zhang – Molecular profiling of aromatase
Project 04 – Bates – Neonatal intensive care

If the publication is not available electronically, provide 5 paper copies of the publication.

Note: The grant agreement requires that recipients acknowledge the Pennsylvania Department of Health funding in all publications. Please ensure that all publications listed acknowledge the Department of Health funding. If a publication does not acknowledge the funding from the Commonwealth, do not list the publication.

Title of Journal Article:	Authors:	Name of Peer-reviewed Publication:	Month and Year Submitted:	Publication Status (check appropriate box below):
1. Prediction of phenotypes of missense mutations in human proteins from the structures of biological assemblies	Wei, Q, Xu, Q, **Dunbrack, RL Jr.**	Proteins: Structure, Function, Bioinformatics	May 2012	☐Submitted ☐Accepted ☒Published
2. The role of balanced training and testing data sets for binary classifiers in bioinformatics.	Wei Q, **Dunbrack RL Jr.**	PLOS ONE	November 2012	☐Submitted ☐Accepted ☒Published
3. Assignment of protein sequences to existing domain and family classification systems: Pfam and the PDB.	Xu, Q and **Dunbrack, RL Jr.**	Bioinformatics	February 2012	☐Submitted ☐Accepted ☒Published

20(B) Based on this project, are you planning to submit articles to peer-reviewed publications in the future?

Yes____X_____ No_____

If yes, please describe your plans:

Several papers are in the revision process and have not yet been submitted. Here is a list of the current complete drafts and where we plan to submit them:

1. Xu et al. Autophosphorylation complexes in crystals of protein kinases. *Science Signaling*
2. Wei and Dunbrack. Structural analysis of Pfam clans. *Bioinformatics*
3. Xu and Dunbrack. Common protein/protein and peptide/protein interfaces in protein crystals on the domain level. *Nature Methods*

21. Changes in Outcome, Impact and Effectiveness Attributable to the Research Project. Describe the outcome, impact, and effectiveness of the research project by summarizing its impact on the incidence of disease, death from disease, stage of disease at time of diagnosis, or other relevant measures of outcome, impact or effectiveness of the research project. If there were no changes, insert "None"; do not use "Not applicable." Responses must be single-spaced below, and no smaller than 12-point type. DO NOT DELETE THESE INSTRUCTIONS. There is no limit to the length of your response.

None.

22. Major Discoveries, New Drugs, and New Approaches for Prevention Diagnosis and Treatment. Describe major discoveries, new drugs, and new approaches for prevention, diagnosis and treatment that are attributable to the completed research project. If there were no major discoveries, drugs or approaches, insert "None"; do not use "Not applicable." Responses must be single-spaced below, and no smaller than 12-point type. DO NOT DELETE THESE INSTRUCTIONS. There is no limit to the length of your response.

None.

23. Inventions, Patents and Commercial Development Opportunities.

23(A) Were any inventions, which may be patentable or otherwise protectable under Title 35 of the United States Code, conceived or first actually reduced to practice in the performance of work under this health research grant? Yes_____ No___X_____

If "Yes" to 23(A), complete items a – g below for each invention. (Do NOT complete items a - g if 23(A) is "No.")

a. Title of Invention:

b. Name of Inventor(s):

c. Technical Description of Invention (describe nature, purpose, operation and physical, chemical, biological or electrical characteristics of the invention):

d. Was a patent filed for the invention conceived or first actually reduced to practice in the performance of work under this health research grant?
Yes_____ No____

If yes, indicate date patent was filed:

e. Was a patent issued for the invention conceived or first actually reduced to practice in the performance of work under this health research grant?
Yes_____ No____
If yes, indicate number of patent, title and date issued:
Patent number:
Title of patent:
Date issued:

f. Were any licenses granted for the patent obtained as a result of work performed under this health research grant? Yes_____ No____

If yes, how many licenses were granted?_____

g. Were any commercial development activities taken to develop the invention into a commercial product or service for manufacture or sale? Yes___ No____

If yes, describe the commercial development activities:

23(B) Based on the results of this project, are you planning to file for any licenses or patents, or undertake any commercial development opportunities in the future?

Yes_____ No____X_____

If yes, please describe your plans:

24. Key Investigator Qualifications. Briefly describe the education, research interests and experience and professional commitments of the Principal Investigator and all other key investigators. In place of narrative you may insert the NIH biosketch form here; however, please limit each biosketch to 1-2 pages.

BIOGRAPHICAL SKETCH

Provide the following information for the Senior/key personnel and other significant contributors.
Follow this format for each person. **DO NOT EXCEED FOUR PAGES.**

NAME	POSITION TITLE
Roland L. Dunbrack, Jr., Ph.D.	Professor
eRA COMMONS USER NAME (credential, e.g., agency login) DUNBRACK	

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	MM/YY	FIELD OF STUDY
Harvard College, Cambridge, MA	A.B.	06/85	Chemistry
University of Cambridge, Cambridge, UK		05/87	Theoretical Chemistry
Harvard University, Cambridge, MA	Ph.D.	06/93	Biophysics
University of California, San Francisco, CA	Postdoc	09/97	Pharmaceutical Chem.

A. PERSONAL STATEMENT

My research is concentrated on the development of methods for structure prediction of proteins and protein complexes. Our aims are to develop methods and software to produce models of proteins in biologically relevant states, and to apply these methods to systems of biological and clinical interest. We have collaborated with nearly all of my scientific colleagues at the Fox Chase Cancer Center in applying these methods to understand existing experimental data and to design new experiments aimed at testing hypotheses in various biological processes.

B. POSITIONS

Graduate student, Harvard University, Program in Biophysics, Advisors: Martin Karplus and Jack Strominger	1987-1993
Postdoctoral Fellow, University of California, San Francisco, CA	1993-1997
Assistant Professor, Fox Chase Cancer Center, Philadelphia, PA	1997-2003
Adjunct Assistant/Associate Professor/Professor, Biochemistry and Biophysics, University of Pennsylvania School of Medicine, Philadelphia, PA	2000-present
Director, Organic Synthesis Facility, Fox Chase Cancer Center, Philadelphia, PA	2003-present
Director, Molecular Modeling Facility, Fox Chase Cancer Center, Philadelphia, PA	2003-present
Associate Professor, Fox Chase Cancer Center, Philadelphia, PA	2003-2011
Adjunct Associate Professor, Biochemistry, Drexel University College of Medicine, Philadelphia, PA	2004-present
Member, Fox Chase Cancer Center Keystone Program in Head and Neck Cancer	2008-present
Professor, Fox Chase Cancer Center, Philadelphia, PA	2011-present

MEMBERSHIPS

Editorial Boards – PLOS Computational Biology
Advisory Boards – RCSB Protein Data Bank; Seattle Structural Genomics Center for Infections Disease
NIH Study Sections – NIH MSF-D Chartered Member (2009-2011); total of 29 NIH panels (2001-2013)

NSF Site Visits – Protein Data Bank (May 2003); ITR Biogeometry Project, UNC/Duke (November 2002)

HONORS

Herchel Smith Fellowship, Awarded by Harvard College for 1-5 years study, Emmanuel College, University of Cambridge	1985-1987
Phi Beta Kappa, Harvard College, *Alpha* Chapter of Massachusetts	1985
National Science Foundation Predoctoral Fellowship	1987-1990
National Institutes of Health Postdoctoral Fellowship	1993-1996
Senior Research Excellence Award, Temple University School of Medicine	2012
Special Contributor Award, Fox Chase Cancer Center	2014

C. PUBLICATIONS (not included due to 2 page space limitation)

D. RESEARCH SUPPORT

Ongoing Research Support

R01 GM084453 (PI: Dunbrack) 5/15/2008 – 4/30/2017
NIH Role: Principal Investigator

Bayesian Statistics and Algorithms for Homology Modeling

The major goals of this project are: 1) Non-parametric regression of bond angles in protein side chains and modeling of structural heterogeneity of side chains; 2) Automatic protein family and superfamily assignments of proteins in the PDB and web-based homology modeling of protein complexes; 3) Identification of kinase autophosphorylation complexes in protein crystals and experimental testing of proposed complexes.

R01 GM078221 (PI: Gray, JHU) 9/1/2012 - 8/31/2016
NIH Role: Site PI

Prediction of the Structure of Therapeutic Antibodies with their Antigens

The major goals of this project are to develop new methods for predicting the structures of antibodies and for improved methods of docking antibodies to antigens. My role in this project is to supervise the efforts of graduate student Brian Weitzner in structural bioinformatics analysis of antibody CDRs and the role of beta turns in loop structure prediction.

1/1/2014 - 12/31/2014
Rosetta Commons Role: Principal Investigator

Development and implementation of statistical scoring functions in Rosetta

The major goals of this project are to develop and test improved statistical scoring functions within Rosetta, including backbone-dependent dihedral and bon angle scoring terms for the backbone and side chains.

R03 CA167264 (PI: Egleston) 3/1/2013 - 2/28/2015
NIH Role: Co-Investigator

Clinical Trials with Exclusions Based on Race, Ethnicity, and English Fluency

The major goals of this project are: 1) To identify and describe characteristics of clinical trials that have explicit inclusion or exclusion criteria related to race or ethnicity; and 2) To identify and describe characteristics of clinical trials that have explicit inclusion or exclusion criteria related to language.