MCP Papers in Press. Published on May 18, 2008 as Manuscript M700550-MCP200

MCP Submitted

Targeting the Human Cancer Pathway Protein Interaction Network by Structural Genomics

Yuanpeng Janet Huang[1,], Dehua Hang[1], Long Jason Lu[2], Liang Tong[4], Mark B. Gerstein[2,3], and Gaetano T. Montelione[1,*]

[1]Department of Molecular Biology and Biochemistry, Center for Advanced Biotechnology and Medicine, and Northeast Structural Genomics Consortium, Rutgers University, Piscataway, NJ 08854

[2]Department of Molecular Biophysics & Biochemistry

[3]Department of Computer Science, and Northeast Structural Genomics Consortium Yale University, New Haven, CT 06520

[4]Department of Biological Sciences and Northeast Structural Genomics Consortium, Columbia University, New York, NY 10027

*To whom correspondence should be addressed.

Email: guy@cabm.rutgers.edu

Running title: Human Cancer Pathway Protein Interaction Network

Summary

Structural genomics provides an important approach for characterizing and understanding systems biology. As a step towards better integrating protein three-dimensional (3D) structural information in cancer systems biology, we have constructed a Human Cancer Pathway Protein Interaction Network (HCPIN) by analysis of several classical cancer-associated signaling pathways and their physical protein-protein interactions. Many well-known cancer-associated proteins play central roles as "hubs" or "bottlenecks" in the HCPIN. At least half of HCPIN proteins are either directly associated with or interact with multiple signaling pathways. While some 45% of residues in these proteins are in sequence segments that meet criteria sufficient for approximate homology modeling (Blast E-val < 10^{-6}), only ~ 20% of residues in these proteins are structurally covered using high-accuracy homology modeling criteria (i.e. Blast E_val < 10^{-6} and at least 80% sequence identity) or by actual experimental structures. The HCPIN website (http://nmr.cabm.rutgers.edu/hcpin) provides a comprehensive description of this biomedical important multi-pathway network, together with experimental and homology models of HCPIN proteins useful for cancer biology research. In order to complement and enrich cancer systems biology, the Northeast Structural Genomics Consortium (NESG) (www.nesg.org) is targeting > 1,000 human proteins and protein domains from the HCPIN for sample production and 3D structure determination. The long-range goal of this effort is to provide a comprehensive 3D structure-function database for human cancer-associated proteins and protein complexes, in the context of their interaction networks. The network-based target selection (BioNet) approach described here is an example of a general strategy for targeting co-functioning proteins by structural genomics projects.

Introduction

In the past decades, many cancer-associated genes have been discovered, their mutations precisely identified, and the pathways through which they act characterized (1-3). The completion of the human genome sequence (4-6), the use of automated sequencing technology, and the development of microarray-based genomics and proteomics technologies (7,8), have had a significant impact on the field of cancer biology (9). In part based on these genome-scale data, cancer is now recognized as a systems biology disease (10). Accordingly, a comprehensive analysis of the molecular basis of cancer requires integration of the distinct, but complementary fields of biochemistry, genomics, cell biology, proteomics, structural biology, and systems biology (8).

Recently, a large number of biological pathway and network databases have been developed to capture the expanding knowledge of protein-protein interactions (e.g., HPRD (11) and DIP (12)) and of metabolic and/or signaling pathways (e.g., KEGG (13), Reactome (14), STKE – http://stke.sciencemag.org, and BioCarta - http://www.biocarta.com). A few databases are specifically focused on cancer-associated signaling pathways, such as The Cancer Cell Map (http://cancer.cellmap.org) and the Rel/NF-κB Signal Transduction Pathway (http://www.nf-kb.org). Pathguide (15) provides an overview of more than 200 web-based biological pathway and network databases. It is challenging to appropriately integrate and utilize this large number of individual databases for systems biology (16). Lu *et al.* (17) have proposed to merge both pathway and network approaches by embedding pathways into large-scale network databases. This approach integrates data on classical biochemical pathways with newly generated large scale proteomics data.

Since the era of genome sequencing, biologists have made extensive use of protein sequence information. 3D structural information is increasingly being used for understanding evolution and the mechanisms of molecular function. Three-dimensional (3D) structure provides critical information connecting protein sequence with molecular function. While sequence alignments, which are broadly used by the molecular biology community, provide useful suggestions about which residues in homologous protein sequences are in corresponding positions, 3D structure-based alignments provide the true determination of corresponding residue positions (18-20), which may be inaccurately identified by sequence alignment information alone especially in cases where the sequence conservation is weak. In favorable cases, protein structure can yield insights into mechanisms of enzyme activities and protein-ligand interactions. In addition, 3D structures of proteins involved in human disease can be used to discover and/or optimize new pharmaceutical agents (21,22).

A complete understanding of molecular interactions requires high-resolution 3D structures, as they provide key atomic details about binding interfaces and information about structural changes that accompany protein-protein interactions. Structural genomics is an international effort aimed at providing 3D structures, either directly by X-ray crystallography or NMR spectroscopy, or by homology modeling, for all proteins in nature (23). Such a comprehensive structure-function database, containing experimental structures and homology models for hundreds of thousands of proteins, will accelerate research in all areas of biomedicine (24-26).

Recently, Xie and Bourne (27) have discussed the structural coverage of human proteins grouped by the Enzyme Commission and the Gene Ontology classifications. This analysis provides a valuable summary of the structural information available for many human disease-related proteins, and provides guidance for protein target selection by structural genomics projects.

As a component of this vision of structural genomics, we have established the Human Cancer Pathway Protein-Interaction Network (HCPIN) database, a collection of human proteins that participate in cancer-associated signaling pathways, and their protein-protein interactions. HCPIN (version 1.0) includes ~3000 proteins and ~10,000 interactions. HCPIN integrates (embeds) pathway data with protein-protein interaction data (17), and provides protein structure-function annotations to inform cancer biology. The HCPIN website (http://nmr.cabm.rutgers.edu/hcpin), illustrated in Fig. 1, provides an extensive collection of experimental and homology models of proteins or domains associated with human cancers.

In this paper we summarize the current 3D structural coverage of HCPIN, and present plans for targeting the remaining proteins in this network for structural analysis. The Northeast Structural Genomics Consortium (NESG) has selected proteins from HCPIN for cloning, expression, purification, and 3D structure determination. This network-based target selection approach provides a framework not only for completing structural coverage of a disease-associated protein interaction network, but also provides specific hypotheses regarding protein interaction partners which can be tested by co-expression, co-crystallization, and 3D structure determination of the resulting protein-protein complexes (28,29). The long-range goal of this

effort is to provide a comprehensive 3D structure-function database for human cancer-associated proteins, the corresponding protein-protein complexes, and their interaction network.

Experimental Procedures

Database Searches

Cell cycle progression, apoptosis, MAPK, Toll-like receptor, TGF-beta, PI3K, and JAK-STAT signal transduction pathways were downloaded from KEGG database (version 0.6, January 2006) (13). Protein-protein interactions, and multi-protein complexes were downloaded from Human Protein Reference Database (11) (09_13_05 release), which included ~16,000 proteins and ~20,000 interactions. Interactions for all pathway proteins and also additional interactions between interaction proteins are included in the HCPIN network. The list of 363 genes involved in human cancer was obtained from the Cancer Gene Census (CGC) Database (http://www.sanger.ac.uk/genetics/CGP/Census) (1). This list is exclusively restricted to genes in which mutations that are reported are causally implicated in oncogenesis. We used IPI human cross reference file (release 3.12) (30) to cross reference proteins from HCPIN, CGC, and SwissProt (31).

HCPIN 3D structural coverage statistics is assessed by running a BLAST search against PDB sequences (February, 2006), using the TargetDB search tool (http://targetdb.pdb.org/) with standard default parameters. Disordered residues, with missing coordinates for segments within otherwise well-determined 3D structures, are counted as "structurally covered" in our structural coverage statistics. HCPIN proteins with no cross-referenced SwissProt ID are considered as not having verified gene models, and are excluded from structure statistical analysis.

Bioinformatics Programs

SignalP v3.0 (32) and TMHMM v2.0 (33) were used for predicting secreted and trans-membrane proteins. The Pfam domains are identified in the Swisspfam file provided from Pfam v19.0 (34,35). The program COILS (36) was used to predict coiled-coil regions. We labeled regions of low complexity by using the program SEG (37). Default options were used for all programs. A in-house Perl program was written to predict disorder regions based on mean charge and mean hydrophobicity (38).

Topology and Statistics Analysis

Program pajek is used for network topology analysis (39). The program R was used for statistics analysis (40).

Homology Modeling and Structure Quality Assessment

HCPIN homology models are selected from MODBASE (41) and/or built using the XPLOR homology modeling protocol of HOMA (42). If multiple models are available from MODBASE, the model with highest sequence identity is selected by HCPIN. Structure quality reports for each of the experimental structures and models were generated using the Protein Structure Validation Software suite (43), which includes structure validation analysis with ProsaII (44), Verify3D (45), Procheck (46), MolProbity (47), and other structure quality assessment tools. Over time, the homoly model database of HCPIN will be updated and expanded.

The HCPIN Web-Accessed Database

Generation of web pages (HTML) for the HCPIN server was done using JAVA and a relational database (MySQL). We recommend the following web browsers: Firefox version 2.0 or higher, and Internet Explorer 7 or higher, to provide full JAVA functionality. Ribbon diagrams were generated using PyMOL. We plan to update structure coverage annotation information weekly and update HCPIN protein information every four months.

Results

Human Cancer Pathway Protein-Interaction Network

The Human Cancer Pathway Protein-Interaction Network (HCPIN) is a collection of proteins from cancer-associated signaling pathways together with their protein-protein interactions. The HCPIN version 1.0 was constructed by combining proteins from seven KEGG (13) classical cancer-associated signaling pathways, together with protein-protein interaction data from the Human Protein Reference Database (HPRD) (11). HPRD is a resource of protein-protein interaction information manually collected from the literature and curated by expert biologists to reduce errors (11). We used KEGG because of its high quality (48). Pathway interaction information from KEGG was excluded from HCPIN, owing to lack of precise definitions (17).

The seven pathways in this initial version of HCPIN include (i) cell cycle progression, (ii) apoptosis, (iii) MAP kinase (MAPK), (iv) innate immune response (Toll-like receptor), (v) TGF-beta, (vi) phosphotidyl inositol kinase (PI3K), and (vii) JAK-STAT pathways. Many well-known important cancer-associated proteins, such as P53 and NF-κB, are associated with at least one of these pathways. The current version of HCPIN includes 2977 proteins and 9784 protein-protein

interactions, including 240 multiprotein complexes each comprised of at least 3 proteins (Table 1).

HCPIN proteins collected from the KEGG pathways are called *pathway proteins*. Other HCPIN proteins that are not included in the KEGG pathways but interact with these pathway proteins are called *interaction proteins*. The representation of protein complexes using binary protein-protein interaction graph remains a challenge, as without detailed structural studies it is often not possible to distinguish direct physical interactions from interactions mediated through the complex (49,50). We used triangular pseudonodes, which link proteins involved in the same complex, to represent multi-protein complexes (50). These multi-protein complexes account for ~1000 edges, of the total ~ 10,580 edges, in the HCPIN network. Table 1 summarizes other statistics of the HCPIN network with and without these multi-protein complexes. Of 664 pathway proteins defined by KEGG, 150 have no annotated physical interactions in HPRD. Some of these may be associated with the seven KEGG pathways by gene transcription, or have interaction partners that are not yet identified or annotated in the HPRD database.

The interaction data included in the current version of HCPIN is a subset of HPRD. Although including only ~15% of HPRD proteins, HCPIN accounts for about half of the protein-protein interactions in the HPRD database (09_13_05 release). Despite the fact that HCPIN represents only a portion of signaling network of the human interactome, its degree distribution is similar to that of many other scale-free interactome networks (51-56). The clustering coefficient in the HCPIN network is better approximated by $C(k) \propto k^{-1}$ than by a k-independent clustering coefficient C(k), which further indicates HCPIN's modularity (57,51,52). Future

expansions and refinements of HCPIN will include cancer-related signaling pathways from other sources (15), as well as protein-protein interaction data from other manually-curated sources (e.g., DIP (12), MINT (58), or Reactome (14)). We envision HCPIN as an evolving, curated resource of structure-function information for the human cancer protein interactome.

The Cancer Gene Census Database comprises 363 protein-encoding human genes that are causally implicated in oncogenesis (1), defined here as CGC proteins. Among these 363 CGC proteins, 186 CGC proteins are included in the HCPIN network, and only 52 of these are pathway proteins. This high coverage of cancer genes in the HCPIN confirms that the cancer genes are heavily associated with signaling pathways and their interactions, and also demonstrates that the 7 pathways that we selected for this analysis are central to cancer biology. This coverage may be increased by including additional cancer-related signaling pathways. Many HCPIN proteins that are fundamental in cancer biology, such as Grb2, Jun, Src, etc, are not included in CGC, and many CGC proteins are not included in HCPIN because they are not characterized to date in the protein-protein interaction literature covered by KEGG or HPRD.

Network centrality measures vs. essentiality

The *degree* of a protein (node) is defined as the number of interactions a particular protein participates in (vertex degree). The *betweenness* of a protein (vertex betweenness) measures the number of non-redundant shortest paths going through this protein. Proteins with high degree or high betweenness are *central proteins*, which are often critical for cell survival (59-63). For many scale-free interaction networks, degree and betweenness are highly correlated (63). Similarly, strong correlations are observed for the HCPIN (Kendall's $\tau = 0.79$, P-value < 2.2e-

16). As can be seen in Fig. 2, top central proteins of HCPIN with both high degree and high betweenness include key cancer-associated essential proteins, such as P53, Grb2, Raf1, EGF receptor (EgfR), and others. Fig. 2 also shows that proteins with high betweenness but low degree are quite abundant, especially for CGC proteins (in red). This suggests that bottleneck proteins, like hub proteins, play essential biological roles, which is in agreement with previous observations (61-63).

Crosstalk between signaling pathways

Signaling pathways interact with one another to form complex networks (64). The sub-network of proteins in a specific pathway together with their interaction partners forms a *pathway interaction subnet* (also called embedded pathways (17)). Accordingly, the seven core KEGG signaling pathways used to construct this version of HCPIN are associated with seven larger pathway interaction subnets. We have also estimated here the crosstalk of the seven signaling pathways by looking at the frequencies of specific proteins in (i) each of the seven signaling pathways, and (ii) in each of the seven associated pathway-interaction subnets.

We first analyzed the crosstalk between pathway proteins associated with each of the seven KEGG signaling pathways. About 20% of all HCPIN pathway proteins are included in more than one KEGG signaling pathway. Fig. 3A summarizes the frequency of observing one pathway protein in multiple signaling pathways. For example, the AKT family of paralogs, the phosphoinositide 3-kinase (PI3K) family of paralogs, and the TNFα protein are involved in four of the seven signaling pathways. The uniqueness of particular proteins to particular KEGG pathways differs for the different signaling pathways. While some 60-70% of pathway proteins

from either the innate immune response and apoptosis pathways are directly associated with at least one other signaling pathway, for the other pathways studied only ~30% of pathway proteins are associated with more than one pathway.

We next analyzed the crosstalk between the *pathway interaction subnets* associated with each of the seven KEGG signaling pathways by HPRD interaction data. Fig. 3B summarizes the frequency of observing one HCPIN protein in multiple pathway interaction subnets. These data show that HCPIN proteins are frequently shared between multiple pathway interaction subnets. Overall, about 53% of HCPIN proteins are associated with more than one pathway interaction subnet. In other words, more than half of HCPIN proteins are either directly associated with, or interact with, multiple signaling pathways. Although only ~ 20% of all pathway proteins are directly associated with multiple (> 1) pathways (Fig. 3A), ~ 58% of pathway proteins are associated with multiple pathway interaction subnets (Fig. 3C). The percentage of pathway proteins associated with multiple pathway interaction subnets (58%) is similar to the percentage of all HCPIN proteins associated with these interaction subnets (53%); the cross talk between pathways is mediated approximately equally by core pathway proteins and interaction proteins.

Seven *pathway proteins* are involved in all seven pathway-interaction subnets (i.e., Raf1 – a serine/threonine kinase, Stat1, Stat3, Rb, P53, CBP, TGFR1). Another seven *interaction proteins* (i.e. proteins in the interaction subnet that are not core pathway proteins) are included in all seven pathway-interaction subnets (i.e. tyrosine kinase Lyn, estrogen receptor alpha, β catenin, insulin receptor, casein kinase II, Hsp90-alpha, and Sam68). These proteins associated with all seven interaction subnets play central roles in cancer biology.

Structural Coverage of HCPIN Proteins

The accuracy of homology models is largely determined by the percent sequence identity with the template 3D structure upon which the model is based (65,43). Models built at ~30-50% sequence identity with the template (a medium-accuracy modeling level) tend to have ~90% of the main-chain modeled within 1.5 Å RMS deviations from the correct structure, but with frequent side-chain packing, core distortion, and loop conformation errors (65,66). Homology models built with more than 50% sequence identity tend to have about 1.0 Å RMS deviation from correct structures for the main-chain atoms, with larger deviations for side-chain packing (66). Our goal is to characterize the structural coverage of the HCPIN using high quality experimental structures or accurate models, especially for enzyme active sites, based on structural templates with Blast E_val < 10^{-6} and sequence identity > 80% (a high-accuracy modeling level). Although this cutoff is somewhat arbitrary, models generated from such templates will usually be of high reliability and accuracy. Such high-quality structures or models of these human proteins are potentially useful for active site docking, studying catalytic mechanism, and designing ligands useful for drug discovery (67).

We have estimated the structural coverage of HCPIN at both *medium-accuracy modeling level* (defined here as Blast E value < 10^{-6}), and *high-accuracy modeling level* (defined here as Blast E value < 10^{-6} and at least 80% sequence identity). Human protein sequence information has been annotated by different experimental and computational methods, and stored in different databases with various levels of gene model accuracy (30). Alternative splice sites, translation initiation sites, and other gene modeling issues complicate the protein sequence annotation

process (68). SwissProt is a high-quality manually annotated protein knowledgebase (69). About 78% of HCPIN protein sequences (2328 sequences) can be validated by SwissProt (IPI v3.12) gene model annotations (30). The structural coverage statistics discussed here are for only these 2328 protein sequences that can be verified by SwissProt data.

Table 2 summarizes the structural coverage of HCPIN proteins at medium- and high-accuracy homology modeling levels. At medium-accuracy level, about 86% of SwissProt verified proteins from the seven HCPIN pathways (pathway proteins) have at least one domain with structural information available from the PDB. These proteins are defined as having *single-domain coverage* (27); i.e. either an experimental structure or a structure template useful for medium-accuracy modeling of at least part of the protein structure. These structures and models cover about 55% of residues in HCPIN (define here as *residue coverage*), excluding predicted low-complexity and coiled-coil regions. Interestingly, innate immune response and apoptosis pathways, which are heavily involved in pathway cross talk, also have the highest residue coverage (>70%). At the high-accuracy modeling level, the structural coverage of pathway proteins is much lower; only 52% have single-domain coverage, with 25% of residues covered. These structural coverage statistics are upper bounds, since this analysis excludes the ~20% of proteins in HCPIN for which protein coding sequences cannot be verified by the SwissProt (IPI v3.12) database.

The single-domain and residue coverage of the *interaction proteins*, which are included in the seven pathway-interaction subnets but not in the seven KEGG pathways, is much lower than for pathway proteins; 76% and 42%, respectively, at medium-accuracy level and 44% and

18%, respectively, at high-accuracy level. These coverage statistics reflect the traditional bias of targeting core signaling pathway proteins in structural biology projects. Overall, HCPIN has 78% (45%), 46% (20%) single-domain (residue) coverage, at medium- and high- accuracy modeling levels, respectively. This single-domain coverage of HCPIN proteins is significantly higher than the estimated average single-domain coverage of the human proteome (27).

We have annotated the 3D structural coverage of all HCPIN proteins in the network diagrams provided on the HCPIN web site (Fig 1B). These web-based graph representations provide direct interactive global views of the 3D structural coverage of these pathway protein interaction networks. The outside ring on each node represents the percentage of protein's residue coverage.

HCPIN domains

Domains are the evolutionary modular building blocks of proteins. Experimental protein structure determination processes using X-ray crystallography or NMR spectroscopy are generally domain oriented. Pfam is a manually curated database of protein domain families derived from sequenced genomes (34). There are ~ 1000 PfamA domains identified in HCPIN, with size ranging from ~ 50 to 1000 residues. At medium-level modeling accuracy, 53% of HCPIN PfamA domain families have complete *fold coverage* (i.e. at least one member of the domain family has essentially complete 3D structural coverage), while 35% of HCPIN domain families have no fold coverage at all. About 10% of Pfam domain families in HCPIN have partial fold coverage; i.e. a 3D structure is available for part of the predicted Pfam domain. This

reflects inherent differences between sequence-alignment based domain boundaries used in Pfam and the actual structural domain boundaries.

Some 10% of HCPIN domains occur at least 10 times in the set of HCPIN proteins. The most abundant domain in HCPIN is the Collagen domain (appearing 265 times), which occurs frequently in extracellular structural proteins involved in formation of connective tissue. Other frequently occurring domain types include Pkinase, zf-C2H2, and WD40 domains.

Table 3 summarizes the top 2% most abundant domain types in the HCPIN together with their 3D structural coverage statistics. All 21 of these most abundant domain families have 'complete fold coverage', in that at least a medium-level accuracy model or experimental structure is available for the full sequence of the domain. *Modeling coverage*, defined here as the percentage of domain members in HCPIN that can be modeled at high-level modeling accuracy, is also summarized for these domain families in Table 3. The frequently occurring domain families of intracellular proteins listed in Table 3 have relatively high modeling coverage. For example, the SH2, an intracellular signaling domain, has the highest modeling coverage, 58%. Experimental structures are available for fewer members of the frequently occurring secreted and membrane-associated domains listed in Table 3, resulting lower modeling coverage of these domain families. Progress in completing the HCPIN modeling coverage for these most abundant domains of HCPIN will provide a comprehensive understanding across the domain family of their structure – function relationships.

The HCPIN structure gallery

The HCPIN web site includes over 1000 protein or domain structure models, of which two-thirds are experimental structures from the PDB (with greater than 99% sequence identity to the human HCPIN protein or protein domain), and one-third are homology models built with structural templates having Blast E-value < 10^{-6} and at least 80% sequence identities (Fig. 1D). To date, the NESG structural genomics project has determined 3D structure of ten human proteins or domains targeted from the HCPIN; some of these are shown in Fig 4.

HCPIN Target Selection for Structural Genomics

With the goal of providing high-accuracy structural models of disease-associated human proteins, especially enzymes, our homology models of HCPIN proteins require a template protein of known 3D structure with pairwise Blast E-value < 10^{-6} and > 80% sequence identity with the target protein (67). As discussed above, our structure coverage analysis shows that significant experimental efforts in X-ray crystallography and/or NMR spectroscopy are still needed to complete the structure coverage of the HCPIN network at this high-accuracy modeling level. Accordingly, these "structurally-uncovered" regions (defined at this high-accuracy level) of HCPIN proteins have been selected for sample production and structure analysis efforts by the Northeast Structural Genomics Consortium.

Are HCPIN proteins suitable for structural genomics efforts?

Due to limitations of current protein structure production technologies, it is generally more challenging to determine 3D structures of eukaryotic proteins, or of secreted, integral membrane, or one-pass transmembrane proteins, compared with intracellular proteins. Integral membrane

proteins are particularly challenging to produce for 3D structure analysis. About 10% of HCPIN intracellular proteins have 100% residue coverage at high-accuracy level, while only ~2% of HCPIN proteins predicted to be secreted and/or membrane-associated have such complete coverage (Fig. 5A). However, considering the HCPIN proteins with only partial structural coverage, our analysis shows that pathway proteins predicted to be secreted, and/or membrane-associated (e.g. soluble domains of one-pass transmembrane proteins), have similar single-domain and residue coverage compared with intracellular proteins (Table 2). These statistics suggest that structural genomics should not only target domains from intracellular proteins, but also the domain families of extracellular secreted and/or extracellular domains of one-pass trans-membrane human proteins of the HCPIN.

Size limitations are a concern for structural genomics that require large-scale protein sample production, and are particularly relevant for structural NMR studies. Protein sample production is generally more successful for proteins of < 600 residues. NMR studies usually require samples with < 180 residues. For this reason, we also analyzed size distributions of Swiss-Prot validated HCPIN protein chains (Fig. 5). The average full-length HCPIN protein is about 600 residues. The size distribution of predicted intracellular proteins is similar to the size distributions of predicted secreted and membrane-associated proteins (Fig. 5B). Size distributions are also similar for proteins with and without structural single-domain coverage (Fig. 5B). Even very large proteins contain domains with some structural coverage, which in most cases have been studied by expressing segments of the protein sequence constituting one or a few structural domains. Residue structural coverage distributions (Fig. 5C) are also similar for predicted intracellular, secreted and/or membrane-associated HCPIN proteins, with an average

coverage of ~110 residues. These size statistics are within the size limitation ranges that are currently addressed well by structural genomics efforts, supporting the feasibility of including these HCPIN network proteins as targets of the Northeast Structural Genomics Consortium.

Target selection process

Fig. 6 shows details of our target selection process. HCPIN v1.0 consists of 664 pathway proteins identified from KEGG, together with additional 2313 interaction proteins from HPRD. 2328 of these 2977 HCPIN proteins are validated by SwissProt (IPI v3.12)(30). For each amino acid sequence of those validated proteins, we filtered out regions that are not suitable for high-throughput structural genomics efforts, including regions with low complexity, those predicted to be coiled-coils, or those predicted to be largely disordered (38). We have identified 1160 intracellular proteins that have regions/domains suitable for such high-throughput structural genomics efforts. Domains from secreted or membrane proteins have also been targeted as part of technology-development projects, but with lower priorities.

Fig. 5D shows the size distribution of these targeted intracellular proteins. Although the size of the full-length targeted proteins varies, about 75% of targeted regions/domains have less than 300 residues. These protein targets are publicly accessible at http://nmr.cabm.rutgres.edu:9090/PLIMS. Efforts have begun to clone, express, purify and characterize these 1160 human proteins and protein domains. We have prioritized these targets mainly based on high-throughput feasibility, rather than other factors such as molecular and cellular functions. In addition, we prioritize for sample production and structure analysis hub and bottleneck proteins, with high network degree and/or betweenness measures. The network

annotations in the HCPIN database also provide biological and bioinformatics information is also being used on a case-by-case basis to prioritize particular protein targets.

Discussion

Protein sample production concerns

Protein sample production is challenging for the HCPIN proteins for several reasons. Cloning and expression of certain human proteins in *E. coli* can be difficult or impossible. Many of these signaling proteins have multiple domains, evolved to convey biological signals from different inputs, and require reliable techniques for domain parsing. In addition, these cancer-associated signaling networks include significant numbers of proteins with extensive disordered regions, which are inherently challenging for expression, purification, and structure determination (70,71). Large macromolecular complexes not only require larger amounts of material, but also a precise and coordinated assembly of the different subunits, conditions that are often not easy to reproduce *in vitro* (28).

Despite the challenges, there are certain technical advantages of targeting an extensive protein interaction network like the HCPIN. Many proteins that fail expression when produced alone can be expressed, purified and crystallized by co-expressing and co-purifying them with their interacting partner proteins (29). We are taking advantage of this approach with HCPIN targets, as potential partners for co-expression and co-purification are indicated from the network. While cancer-associated signaling networks are likely to include significant numbers of proteins with extensive disordered regions (70,71), such disordered regions may become ordered upon

binding to their protein partners, making the corresponding complexes suitable for high-throughput structural genomics (72-75).

General Strategy for Targeting Proteins from Pathway-based Interaction Networks

We propose a general strategy to select targets from pathway-based interaction networks. This target selection strategy can be applied to any biochemical pathway of interest. Previously reported target selection strategies for structural genomics have focused on family (76,77), whole genome (78-80), pathways (67,25), and complexes (28,29). The target selection strategy we present here combines the selection strategies that have been proposed for structural genomics of biochemical pathways (67,25) and of protein-protein complexes (28,29).

First, lists of proteins involved in a specific biological pathway are collected. These proteins are called pathway proteins. Interaction proteins are then identified, including those that potentially interact directly with any pathway proteins, or contribute to multiprotein complexes formed with pathway proteins. Interactions can be derived from the literature, curated peer-reviewed databases (13,12,58,11,14), high-throughput protein interaction experiments (53-56), and/or from integrated prediction methods (81,82). Gene models for both pathway and interaction proteins are then validated using SwissProt (83). Protein sequences not verified by SwissProt will require further analysis to confirm their authenticity. Regions of proteins with known 3D structure information from PDB are identified. Regions of proteins not covered with 3D structure information and also suitable for high-throughput structural determination are then selected as structural genomics targets, with emphasis on hub and bottleneck proteins. This *BioNet target selection strategy* not only provides a systematic approach for complete structure

coverage for disease-associated pathways (67,25), but also provides a framework for studying protein interactions and complexes (28,29).

Community Outreach

Since the era of genome sequencing, biologists now use protein sequence information extensively. However, the general biological community uses much less structural information . The HCPIN website (http://nmr.cabm.rutgers.edu/hcpin/, Fig. 1) is built to make structural information about cancer-related proteins easily accessible to cancer biologists. Our future plan for HCPIN includes mapping SNP/mutation information, protein-protein interactions, and various structural bioinformatics predictions onto the 3D structures, adding gene ontology and structure-based functional annotation, and incorporating microarray and protein expression data. We envision HCPIN as an evolving, curated resource of structure-function information for the human cancer protein interactome.

Many intermediate results, such as expression constructs and biochemical reagents, generated in these ongoing structural genomics efforts are feely available to the biology community. Our structure-function database can be leveraged by many other related initiatives. For example, the National Cancer Institutes (NCI) Initiative for Chemical Genetics (ICG) aims to systematically identify perturbational small molecules for each cancer-related protein coded in the human genome (84). Our structural genomics efforts on HCPIN will provide biochemical and structural information, as well as key reagents, organized at a single web site, beneficial for such chemical genetics studies (85).

Acknowledgements. We thank Drs. T. Acton, J. Everett, C. Gelinas, A. Rabson, and E. White for helpful discussions and comments on the manuscript. This work was supported by NIGMS grant U54-074958 from the Protein Structure Initiative of the National Institutes of Health.

References

1. Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., and Stratton, M. R. (2004) A census of human cancer genes, *Nat Rev Cancer* **4**, 177-183
2. Vogelstein, B., and Kinzler, K. W. (2004) Cancer genes and the pathways they control, *Nat Med* **10**, 789-799
3. Greenman, C., Stephens, P., Smith, R., Dalgliesh, G. L., Hunter, C., et al. (2007) Patterns of somatic mutation in human cancer genomes, *Nature* **446**, 153-158
4. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., et al. (2001) Initial sequencing and analysis of the human genome, *Nature* **409**, 860-921
5. Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., et al. (2001) The sequence of the human genome, *Science* **291**, 1304-1351
6. Consortium, I. H. G. S. (2004) Finishing the euchromatic sequence of the human genome, *Nature* **431**, 931-945
7. Liang, P., and Pardee, A. B. (2003) Analysing differential gene expression in cancer, *Nat Rev Cancer* **3**, 869-876
8. Wulfkuhle, J., Espina, V., Liotta, L., and Petricoin, E. (2004) Genomic and proteomic technologies for individualisation and improvement of cancer treatment, *Eur J Cancer* **40**, 2623-2632
9. Strausberg, R. L., Simpson, A. J., Old, L. J., and Riggins, G. J. (2004) Oncogenomics and the development of new cancer therapies, *Nature* **429**, 469-474
10. Hornberg, J. J., Bruggeman, F. J., Westerhoff, H. V., and Lankelma, J. (2006) Cancer: a Systems Biology disease, *Biosystems* **83**, 81-90
11. Peri, S., Navarro, J. D., Amanchy, R., Kristiansen, T. Z., Jonnalagadda, C. K., et al. (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans, *Genome Res* **13**, 2363-2371
12. Xenarios, I., Salwinski, L., Duan, X. J., Higney, P., Kim, S. M., and Eisenberg, D. (2002) DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions, *Nucleic Acids Res* **30**, 303-305
13. Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes, *Nucleic Acids Res* **27**, 29-34
14. Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., Jassal, B., Gopinath, G. R., Wu, G. R., Matthews, L., Lewis, S., Birney, E., and Stein, L. (2005) Reactome: a knowledgebase of biological pathways, *Nucleic Acids Res* **33**, D428-432
15. Bader, G. D., Cary, M. P., and Sander, C. (2006) Pathguide: a pathway resource list, *Nucleic Acids Res* **34**, D504-506
16. Cary, M. P., Bader, G. D., and Sander, C. (2005) Pathway information for systems biology, *FEBS Lett* **579**, 1815-1820
17. Lu, L. J., Sboner, A., Huang, Y. J., Lu, H. X., Gianoulis, T. A., Yip, K. Y., Kim, P. M., Montelione, G. T., and Gerstein, M. B. (2007) Comparing classical pathways and modern networks: towards the development of an edge ontology, *Trends Biochem Sci* **32**, 320-331
18. Holm, L., and Sander, C. (1996) Mapping the protein universe, *Science* **273**, 595-603
19. Shindyalov, I. N., and Bourne, P. E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path, *Protein Eng* **11**, 739-747

20. Yang, A. S., and Honig, B. (2000) An integrated approach to the analysis and modeling of protein sequences and structures. I. Protein structural alignment and a quantitative measure for protein structural distance, *J Mol Biol* **301**, 665-678
21. Harris, T. (2000) Genetics, genomics, and drug discovery, *Med Res Rev* **20**, 203-211
22. Weng, Z., and DeLisi, C. (2002) Protein therapeutics: promises and challenges for the 21st century, *Trends Biotechnol* **20**, 29-35
23. Chandonia, J. M., and Brenner, S. E. (2006) The impact of structural genomics: expectations and outcomes, *Science* **311**, 347-351
24. Brenner, S. E. (2001) A tour of structural genomics, *Nat Rev Genet* **2**, 801-809
25. Burley, S. K., and Bonanno, J. B. (2002) Structural genomics of proteins from conserved biochemical pathways and processes, *Curr Opin Struct Biol* **12**, 383-391
26. Mueller, L., and Montelione, G. T. (2002) Structural genomics in pharmaceutical design, *J Struct Funct Genomics* **2**, 67-70
27. Xie, L., and Bourne, P. E. (2005) Functional coverage of the human genome by existing structures, structural genomics targets, and homology models, *PLoS Comput Biol* **1**, e31
28. Bravo, J., and Aloy, P. (2006) Target selection for complex structural genomics, *Curr Opin Struct Biol* **16**, 385-392
29. Strong, M., Sawaya, M. R., Wang, S., Phillips, M., Cascio, D., and Eisenberg, D. (2006) Toward the structural genomics of complexes: crystal structure of a PE/PPE protein complex from Mycobacterium tuberculosis, *Proc Natl Acad Sci U S A* **103**, 8060-8065
30. Kersey, P. J., Duarte, J., Williams, A., Karavidopoulou, Y., Birney, E., and Apweiler, R. (2004) The International Protein Index: an integrated database for proteomics experiments, *Proteomics* **4**, 1985-1988
31. Camon, E., Magrane, M., Barrell, D., Lee, V., Dimmer, E., Maslen, J., Binns, D., Harte, N., Lopez, R., and Apweiler, R. (2004) The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology, *Nucleic Acids Res* **32**, D262-266
32. Bendtsen, J. D., Nielsen, H., von Heijne, G., and Brunak, S. (2004) Improved prediction of signal peptides: SignalP 3.0, *J Mol Biol* **340**, 783-795
33. Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E. L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes, *J Mol Biol* **305**, 567-580
34. Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L., Studholme, D. J., Yeats, C., and Eddy, S. R. (2004) The Pfam protein families database, *Nucleic Acids Res* **32**, D138-141
35. Finn, R. D., Mistry, J., Schuster-Bockler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R., Eddy, S. R., Sonnhammer, E. L., and Bateman, A. (2006) Pfam: clans, web tools and services, *Nucleic Acids Res* **34**, D247-251
36. Lupas, A., Van Dyke, M., and Stock, J. (1991) Predicting coiled coils from protein sequences, *Science* **252**, 1162-1164
37. Wootton, J. C., and Federhen, S. (1996) Analysis of compositionally biased regions in sequence databases, *Methods Enzymol* **266**, 554-571
38. Uversky, V. N., Gillespie, J. R., and Fink, A. L. (2000) Why are "natively unfolded" proteins unstructured under physiologic conditions?, *Proteins* **41**, 415-427
39. Nooy, W. d., Mrvar, A., and Batagelj, V. (2005) *Exploratory social network analysis with Pajek*, Cambridge University Press, Cambridge ; New York

40. Dalgaard, P. (2002) Introductory statistics with R. In., Springer, New York
41. Pieper, U., Eswar, N., Davis, F. P., Braberg, H., Madhusudhan, M. S., Rossi, A., Marti-Renom, M., Karchin, R., Webb, B. M., Eramian, D., Shen, M. Y., Kelly, L., Melo, F., and Sali, A. (2006) MODBASE: a database of annotated comparative protein structure models and associated resources, *Nucleic Acids Res* **34**, D291-295
42. Bhattacharya, A., Wunderlich, Z., Monleon, d., Tejero, R., and Montelione, G. T. (2007) Assessing model accuracy using the homology modeling automatically (HOMA) software, *Proteins*, (in press)
43. Bhattacharya, A., Tejero, R., and Montelione, G. T. (2007) Evaluating protein structures determined by structural genomics consortia, *Proteins* **66**, 778-795
44. Sippl, M. J. (1993) Recognition of errors in three-dimensional structures of proteins, *Proteins* **17**, 355-362
45. Luthy, R., Bowie, J. U., and Eisenberg, D. (1992) Assessment of protein models with three-dimensional profiles, *Nature* **356**, 83-85
46. Laskowski, R. A., MacArthur, M. W., Moss, D. S., and Thornton, J. M. (1993) PROCHECK: a program to check the stereochemical quality of protien structures, *J Appl Crystallogr* **26**, 283-291
47. Lovell, S. C., Davis, I. W., Arendall, W. B., 3rd, de Bakker, P. I., Word, J. M., Prisant, M. G., Richardson, J. S., and Richardson, D. C. (2003) Structure validation by Calpha geometry: phi,psi and Cbeta deviation, *Proteins* **50**, 437-450
48. Wittig, U., and De Beuckelaer, A. (2001) Analysis and comparison of metabolic pathway databases, *Brief Bioinform* **2**, 126-142
49. Ding, C., He, X., Meraz, R. F., and Holbrook, S. R. (2004) A unified representation of multiprotein complex data for modeling interaction networks, *Proteins* **57**, 99-108
50. Ramadan, E., Tarafdar, A., and Pothen, A. (2004) A hypergraph model for the yeast complex network. In. *18th International parallel and distributed processing symposium*, Santa Fe, New Mexico, USA
51. Barabasi, A. L., and Oltvai, Z. N. (2004) Network biology: understanding the cell's functional organization, *Nat Rev Genet* **5**, 101-113
52. Yook, S. H., Oltvai, Z. N., and Barabasi, A. L. (2004) Functional and topological characterization of protein interaction networks, *Proteomics* **4**, 928-942
53. Barrios-Rodiles, M., Brown, K. R., Ozdamar, B., Bose, R., Liu, Z., Donovan, R. S., Shinjo, F., Liu, Y., Dembowy, J., Taylor, I. W., Luga, V., Przulj, N., Robinson, M., Suzuki, H., Hayashizaki, Y., Jurisica, I., and Wrana, J. L. (2005) High-throughput mapping of a dynamic signaling network in mammalian cells, *Science* **307**, 1621-1625
54. Rual, J. F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G. F., Gibbons, F. D., Dreze, M., Ayivi-Guedehoussou, N., Klitgord, N., Simon, C., Boxem, M., Milstein, S., Rosenberg, J., Goldberg, D. S., Zhang, L. V., Wong, S. L., Franklin, G., Li, S., Albala, J. S., Lim, J., Fraughton, C., Llamosas, E., Cevik, S., Bex, C., Lamesch, P., Sikorski, R. S., Vandenhaute, J., Zoghbi, H. Y., Smolyar, A., Bosak, S., Sequerra, R., Doucette-Stamm, L., Cusick, M. E., Hill, D. E., Roth, F. P., and Vidal, M. (2005) Towards a proteome-scale map of the human protein-protein interaction network, *Nature* **437**, 1173-1178
55. Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F. H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S., Timm, J., Mintzlaff, S., Abraham, C., Bock, N., Kietzmann, S., Goedde, A., Toksoz, E., Droege, A., Krobitsch, S.,

Korn, B., Birchmeier, W., Lehrach, H., and Wanker, E. E. (2005) A human protein-protein interaction network: a resource for annotating the proteome, *Cell* **122**, 957-968
56. Lim, J., Hao, T., Shaw, C., Patel, A. J., Szabo, G., Rual, J. F., Fisk, C. J., Li, N., Smolyar, A., Hill, D. E., Barabasi, A. L., Vidal, M., and Zoghbi, H. Y. (2006) A protein-protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration, *Cell* **125**, 801-814
57. Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., and Barabasi, A. L. (2002) Hierarchical organization of modularity in metabolic networks, *Science* **297**, 1551-1555
58. Zanzoni, A., Montecchi-Palazzi, L., Quondam, M., Ausiello, G., Helmer-Citterich, M., and Cesareni, G. (2002) MINT: a Molecular INTeraction database, *FEBS Lett* **513**, 135-140
59. Jeong, H., Mason, S. P., Barabasi, A. L., and Oltvai, Z. N. (2001) Lethality and centrality in protein networks, *Nature* **411**, 41-42
60. Holme, P., Kim, B. J., Yoon, C. N., and Han, S. K. (2002) Attack vulnerability of complex networks, *Phys Rev E Stat Nonlin Soft Matter Phys* **65**, 056109
61. Hahn, M. W., and Kern, A. D. (2005) Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks, *Mol Biol Evol* **22**, 803-806
62. Joy, M. P., Brock, A., Ingber, D. E., and Huang, S. (2005) High-betweenness proteins in the yeast protein interaction network, *J Biomed Biotechnol* **2005**, 96-103
63. Yu, H., Kim, P. M., Sprecher, E., Trifonov, V., and Gerstein, M. (2007) The Importance of Bottlenecks in Protein Networks: Correlation with Gene Essentiality and Expression Dynamics, *PLoS Comput Biol* **3**, e59
64. Weng, G., Bhalla, U. S., and Iyengar, R. (1999) Complexity in biological signaling systems, *Science* **284**, 92-96
65. Marti-Renom, M. A., Stuart, A. C., Fiser, A., Sanchez, R., Melo, F., and Sali, A. (2000) Comparative protein structure modeling of genes and genomes, *Annu Rev Biophys Biomol Struct* **29**, 291-325
66. Baker, D., and Sali, A. (2001) Protein structure prediction and structural genomics, *Science* **294**, 93-96
67. Erlandsen, H., Abola, E. E., and Stevens, R. C. (2000) Combining structural genomics and enzymology: completing the picture in metabolic pathways and enzyme active sites, *Curr Opin Struct Biol* **10**, 719-730
68. Guigo, R., Flicek, P., Abril, J. F., Reymond, A., Lagarde, J., Denoeud, F., Antonarakis, S., Ashburner, M., Bajic, V. B., Birney, E., Castelo, R., Eyras, E., Ucla, C., Gingeras, T. R., Harrow, J., Hubbard, T., Lewis, S. E., and Reese, M. G. (2006) EGASP: the human ENCODE Genome Annotation Assessment Project, *Genome Biol* **7 Suppl 1**, S2 1-31
69. Wu, C. H., Apweiler, R., Bairoch, A., Natale, D. A., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Mazumder, R., O'Donovan, C., Redaschi, N., and Suzek, B. (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information, *Nucleic Acids Res* **34**, D187-191
70. Iakoucheva, L. M., Brown, C. J., Lawson, J. D., Obradovic, Z., and Dunker, A. K. (2002) Intrinsic disorder in cell-signaling and cancer-associated proteins, *J Mol Biol* **323**, 573-584

71. Oldfield, C. J., Ulrich, E. L., Cheng, Y., Dunker, A. K., and Markley, J. L. (2005) Addressing the intrinsic disorder bottleneck in structural proteomics, *Proteins* **59**, 444-453
72. Wright, P. E., and Dyson, H. J. (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm, *J Mol Biol* **293**, 321-331
73. Dunker, A. K., Brown, C. J., Lawson, J. D., Iakoucheva, L. M., and Obradovic, Z. (2002) Intrinsic disorder and protein function, *Biochemistry* **41**, 6573-6582
74. Dyson, H. J., and Wright, P. E. (2002) Coupling of folding and binding for unstructured proteins, *Curr Opin Struct Biol* **12**, 54-60
75. Tompa, P. (2005) The interplay between structure and function in intrinsically unstructured proteins, *FEBS Lett* **579**, 3346-3354
76. Liu, J., Hegyi, H., Acton, T. B., Montelione, G. T., and Rost, B. (2004) Automatic target selection for structural genomics on eukaryotes, *Proteins* **56**, 188-200
77. Chandonia, J. M., and Brenner, S. E. (2005) Implications of structural genomics target selection strategies: Pfam5000, whole genome, and random approaches, *Proteins* **58**, 166-179
78. Kim, S. H. (2000) Structural genomics of microbes: an objective, *Curr Opin Struct Biol* **10**, 380-383
79. Goulding, C. W., Apostol, M., Anderson, D. H., Gill, H. S., Smith, C. V., Kuo, M. R., Yang, J. K., Waldo, G. S., Suh, S. W., Chauhan, R., Kale, A., Bachhawat, N., Mande, S. C., Johnston, J. M., Lott, J. S., Baker, E. N., Arcus, V. L., Leys, D., McLean, K. J., Munro, A. W., Berendzen, J., Sharma, V., Park, M. S., Eisenberg, D., Sacchettini, J., Alber, T., Rupp, B., Jacobs, W., Jr., and Terwilliger, T. C. (2002) The TB structural genomics consortium: providing a structural foundation for drug discovery, *Curr Drug Targets Infect Disord* **2**, 121-141
80. Matte, A., Sivaraman, J., Ekiel, I., Gehring, K., Jia, Z., and Cygler, M. (2003) Contribution of structural genomics to understanding the biology of Escherichia coli, *J Bacteriol* **185**, 3994-4002
81. Brown, K. R., and Jurisica, I. (2005) Online predicted human interaction database, *Bioinformatics* **21**, 2076-2082
82. Rhodes, D. R., Tomlins, S. A., Varambally, S., Mahavisno, V., Barrette, T., Kalyana-Sundaram, S., Ghosh, D., Pandey, A., and Chinnaiyan, A. M. (2005) Probabilistic model of the human protein-protein interaction network, *Nat Biotechnol* **23**, 951-959
83. O'Donovan, C., Martin, M. J., Gattiker, A., Gasteiger, E., Bairoch, A., and Apweiler, R. (2002) High-quality protein knowledge resource: SWISS-PROT and TrEMBL, *Brief Bioinform* **3**, 275-284
84. Tolliday, N., Clemons, P. A., Ferraiolo, P., Koehler, A. N., Lewis, T. A., Li, X., Schreiber, S. L., Gerhard, D. S., and Eliasof, S. (2006) Small molecules, big players: the National Cancer Institute's Initiative for Chemical Genetics, *Cancer Res* **66**, 8935-8942
85. Spring, D. R. (2005) Chemical genetics to chemical genomics: small molecules offer big insights, *Chem Soc Rev* **34**, 472-482

Figure Legends

Figure 1. The Human Cancer Protein Interaction Network (HCPIN) is a web-accessible database. It is designed for use by cancer biologists interested in assessing 3D protein structural information in the context of the protein interaction network. (A) HCPIN home page (http://nmr.cabm.rutgers.edu/hcpin). (B) A snapshot of Networks view, visualizing protein-protein interactions with structure annotations. The outside ring represents the percentage of structural coverage. Green ring – experimental model is available with >99% sequence identities, Yellow ring - homology model is available with >80% sequence identities. The web site provides tools for interactive analysis of the HCPIN network. (C) A snapshot of Proteins view, listing sequence information and PDB blast hits, summarizing all structural information available for the human HCPIN protein and its homologues, and providing links to the corresponding PDB entries and other structure-function annotation information, (D) A snapshot of Icon gallery, a collection of ribbon diagrams for each of the known structures and the structural models in the HCPIN.

Figure 2. Scatter plot of degree and betweenness measures for HCPIN proteins. Black – HCPIN proteins. Red – proteins also listed in Cancer Gene Census Database (1).

Figure 3. Crosstalk between pathways. (A) Frequency of observing one protein in one or more of the 7 KEGG signaling pathways. ~20% of HCPIN pathway proteins are associated with two or more pathways. (B) Frequency of observing one HCPIN protein in one or more of seven pathway interaction subnets. > 50% of HCPIN proteins are associated with two or more interaction subnets. (C) Frequency of observing one pathway protein in one or more pathway interaction subnets. The frequencies (1-7) are also labeled on the side of these pie charts.

Figure 4. Ribbon diagram of some HCPIN proteins/domains solved by NESG. At the bottom of each reprehensive ribbon diagram, we listed Swissprot (SW) name, NESG target id, PDB id, residue coverage and method used for structure determination.

Figure 5. (A) Percent residue coverage distributions for HCPIN proteins, intracellular – proteins inside the cell. s/m – proteins predicted to be secreted or having at least a segment that is integral or trans-membrane. (B) Size distributions of HCPIN proteins and HCPIN proteins with single-domain coverage. intracellular – proteins inside the cell, s/m – as defined above, intracellular-SD – intracellular proteins with single-domain coverage, s/m-SD – proteins predicted to be secreted or having at least a segment that is trans-membrane with single-domain coverage. (C) Size distributions of HCPIN proteins with residue coverage. Intracelluar / s/m-residue - residue coverage of intracellular proteins and predicted secreted/membrane-associated proteins. Single-domain and residue coverages are shown at high-accuracy level. A similar distribution is observed at medium-accuracy level. (D) Boxplots of size distributions of full length and targeted sub-regions of proteins selected by the NESG structural genomics project.

Figure 6. HCPIN target selection process. SEG region: low complexity regions predicted by program SEG (37). SignalP region: signaling peptide predicted by SignalP (32). TM region: transmembrane region predicted by TMHMM (33). C/U-region: structure covered or uncovered

region. T-region: targeted region. Disordered regions are predicted based on mean hydrophobicity and net charge (38) .

Table 1. HCPIN Network Statistics

	HCPIN	HCPIN – pairwise protein-protein interaction only[a]
Proteins/Nodes	2977 proteins (664 pathway proteins) 240 multi-protein complexes	2819 proteins
Interactions	9784[b]	9544
Edges	10583[b] (292 self-interaction loops)	9544 (70 self-interaction loops)
Diameter[c] (longest distance)	11	11
Average distance	4.143	4.086

[a] One multi-protein complex is counted as one interaction, However, it is counted as multiple edges in the HCPIN graph.

[b] Proteins from multi-protein complexes and pseudonodes are excluded for calculation

[c] Measured for the largest connected component

Table 2. Structural Coverage of SwissProt-Validated Proteins from the Seven KEGG Signaling Pathways of the HCPIN

Medium-accuracy homology modeling level
(Blast E_value < 10^{-6})

	Total Structural Coverage			Secreted / Membrane Protein Structural Coverage			Intracellular Protein Structural Coverage		
	No.	%SD[a]	%Res.[b]	%No.[c]	%SD	%Res.	%No.	%SD	%Res.
Apoptosis	72	94	72	19	100	78	81	93	71
TGF	79	90	56	52	90	59	48	89	54
PI3K	81	85	51	12	70	28	88	87	56
Cell-cycle	82	76	36	4	100	31	96	75	36
TLR	88	93	71	45	95	77	55	92	69
JAK-STAT	139	78	54	59	76	49	41	81	58
MAPK	216	94	64	25	98	70	75	92	63
HCPIN									
Pathway proteins	600	86	55	33	85	56	67	86	55
Interaction proteins	1728	76	42	26	75	47	74	76	41
Total	2328	78	45	28	78	49	72	78	44

High-accuracy homology modeling level
(Blast E_value < 10^{-6} and > 80% sequence identity)

	Total Structural Coverage			Secreted /Membrane Protein Structural Coverage			Intracellular Protein Structural Coverage		
	No.	%SD	%Res.	%No.	%SD	%Res.	%No.	%SD	%Res.
Apoptosis	72	68	36	19	64	32	81	69	36
TGF	79	62	29	52	59	26	48	66	32
PI3K	81	36	13	12	10	5	88	39	15
Cell-cycle	82	46	20	4	100	30	96	44	20
TLR	88	68	36	45	75	38	55	63	35
JAK-STAT	139	55	32	59	59	32	41	51	32
MAPK	216	60	33	25	60	36	75	60	32
HCPIN									
Pathway proteins	600	52	25	33	54	26	67	51	25
Interaction protein	1728	44	18	26	39	16	74	46	19
Total	2328	46	20	28	44	18	72	47	20

[a]The percentage of pathway proteins with single-domain structural coverage.

[b]The number of residues covered by PDB hit, divided by total length of proteins in the pathways. Residues predicted to be low complexity or coiled-coil are not counted in the denominator.

[c]The percentage of proteins in the pathways that are predicted to be secreted or have integral or one-pass transmembrane domains.

Table 3. Most Frequently Occurring HCPIN Domains

Pfam Domain Name	Frq	Modeling_coverage [a] (E < 10^{-6} and > 80% seq id)	Molecular Function
Collagen	265	0.01	extracellular structural proteins
Pkinase	184	0.31	protein kinase
zf-C2H2	176	0.17	nucleic acid-binding
WD40	173	0.23	multi-protein complex assemblies
LRR_1	148	0.22	leucine rich repeat, protein-protein interaction
Ank	145	0.40	protein-protein interaction motif
fn3	134	0.18	cell surface binding, signaling
EGF	122	0.12	EGF-like domain
SH3_1	104	0.46	signal transduction related to cytoskeletal organization
Ldl_recept_b	101	0.05	low-density lipoprotein receptor repeakt class B
TPR_1	99	0.32	protein-protein interaction
EGF_CA	94	0.10	calcium binding EGF domain
IQ	81	0.00	calmodulin-binding motif
efhand	79	0.47	calcium-binding domain
ig	77	0.27	immunoglobulin domain
Ldl_recept_a	77	0.13	low-density lipoprotein receptor repeakt class A
SH2	74	0.58	intracellular signaling
Pkinase_Tyr	73	0.41	protein tyrosine kinase
Filamin	72	0.03	actin cross-linking protein
PH	65	0.26	cytoskeleton, intracellular signaling
Spectrin	64	0.08	cytoskeletal structure

[a.] the percentage of domain members that can be modeled at Blast E-val < 10^{-6} and > 80% sequence identity

Figure 1

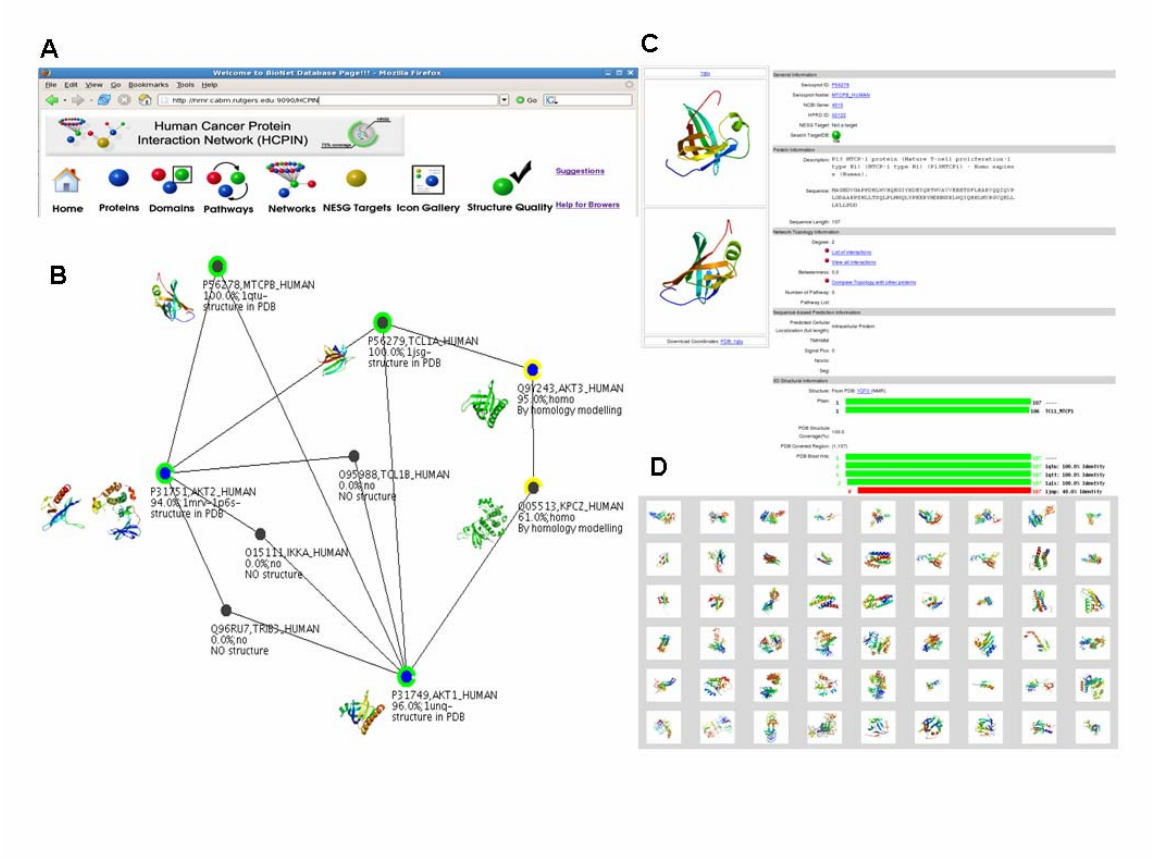


Figure 2

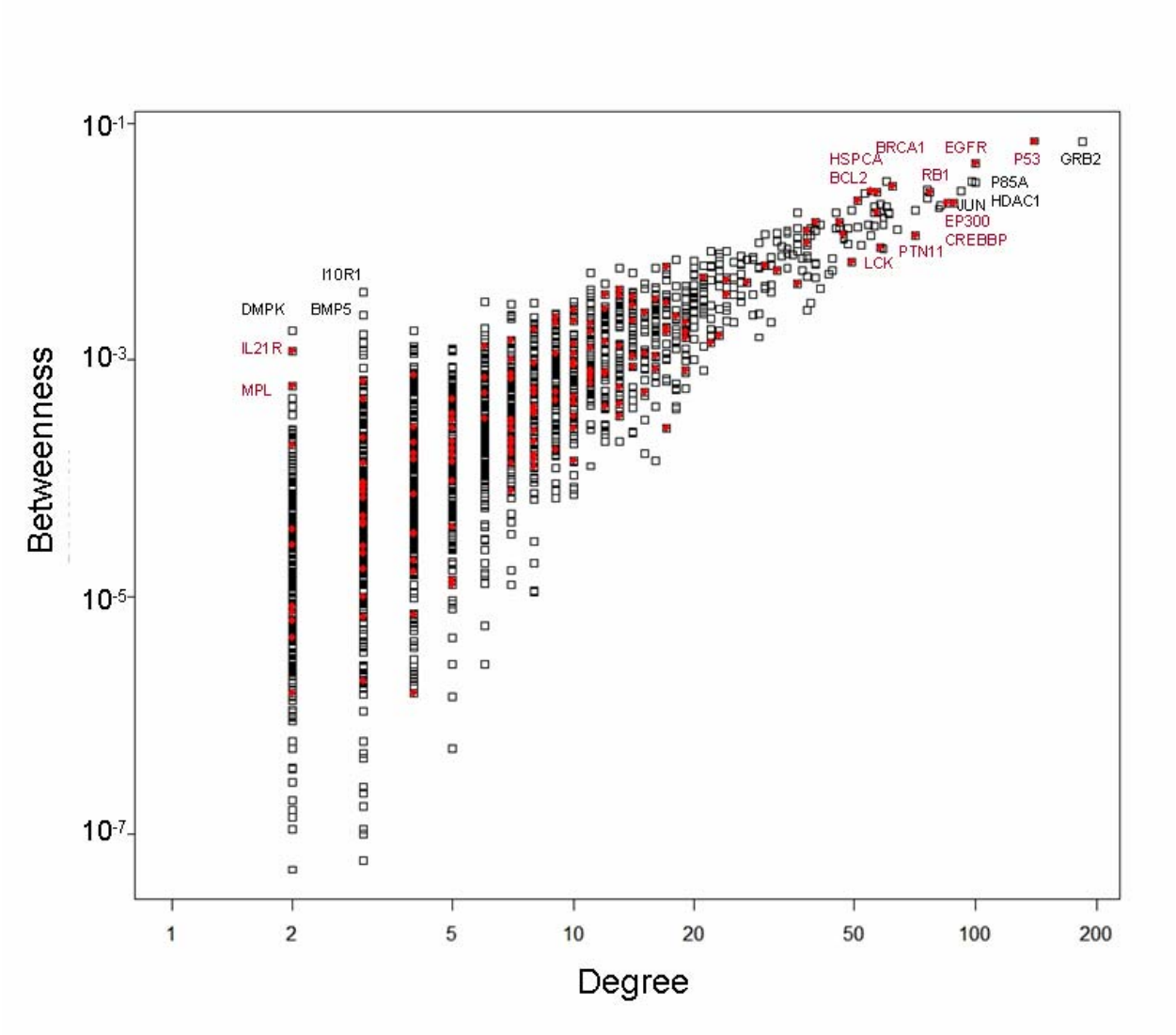


Figure 3

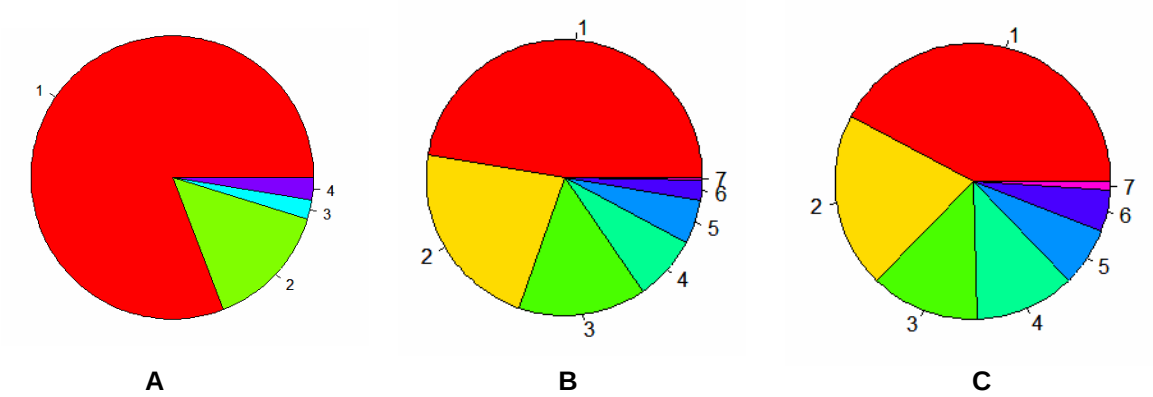


Figure 4

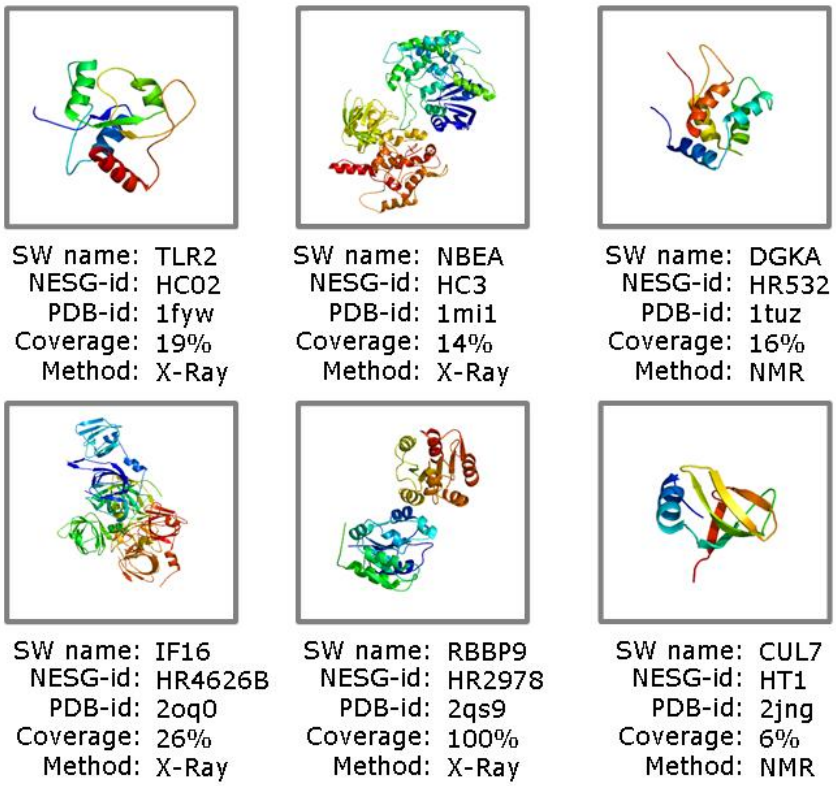
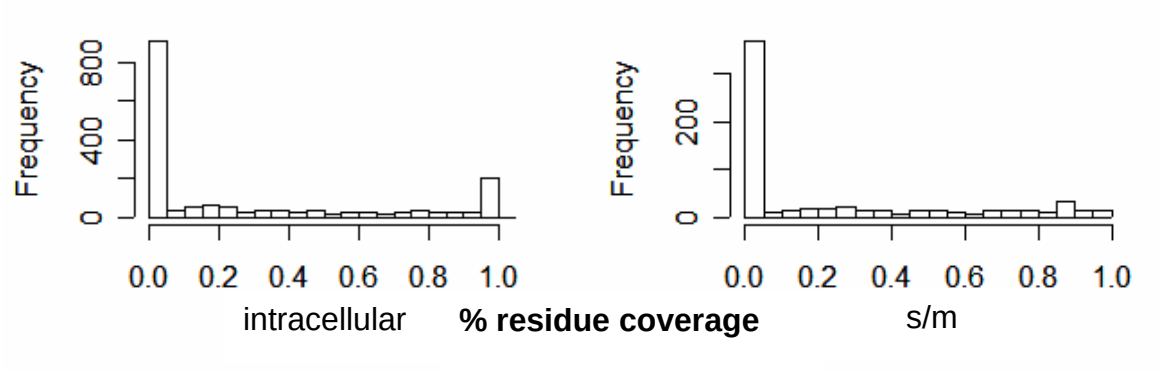
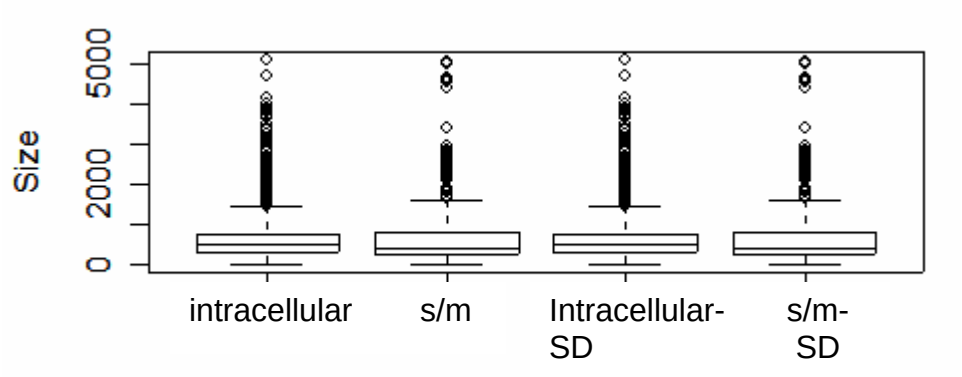


Figure 5

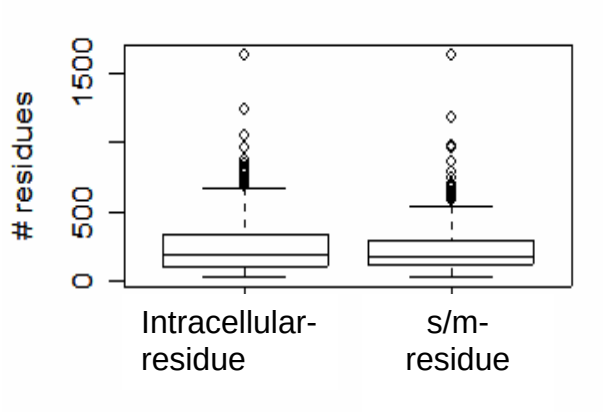
A



B



C



D

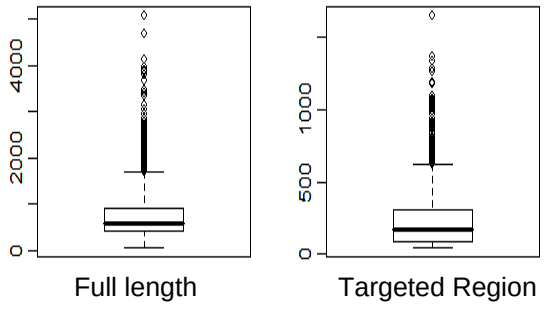


Figure 6