

Computational Tools for the Interactive Exploration of Proteomic and Structural Data*

John H. Morris, Elaine C. Meng, and Thomas E. Ferrin†

Linking proteomics and structural data is critical to our understanding of cellular processes, and interactive exploration of these complementary data sets can be extremely valuable for developing or confirming hypotheses *in silico*. However, few computational tools facilitate linking these types of data interactively. In addition, the tools that do exist are neither well understood nor widely used by the proteomics or structural biology communities. We briefly describe several relevant tools, and then, using three scenarios, we present in depth two tools for the integrated exploration of proteomics and structural data. *Molecular & Cellular Proteomics* 9:1703–1715, 2010.

A 3-D enhanced version of this article is available. The text is identical to this version but includes interactive figures.

Viewing the enhanced version of this article requires the use of a browser plug-in. Please install the plug-in when prompted. <http://www.thesgc.org/iSee/MCP/9/8/e1.html>

MOTIVATION

Structural biology and proteomics provide complementary views of cellular processes. Structural biology is primarily concerned with the structures of biological macromolecules and complexes and the physicochemical interactions they support. Proteomics, on the other hand, tends to take a broader view of how proteins communicate and function within the cell, often encompassing large numbers of proteins that operate in pathways or addressing how groups of proteins work together as a function of time and/or subcellular location. (The term “proteomics,” as used here, includes studies of not only the presence and abundance of proteins under various conditions but also their interactions and their functions, both individually and as parts of larger, more complex systems.) Understanding the molecular interactions between proteins at the atomic level is of obvious utility, yet it is equally critical to understand the broader context of how pathways function and change with differing levels of expression and copy number and how they are controlled by inhibition, activation, and feedback loops.

From the Resource for Biocomputing, Visualization, and Informatics, Department of Pharmaceutical Chemistry, University of California, San Francisco, California 94158-2517

Received, March 16, 2010, and in revised form, June 3, 2010

Published, MCP Papers in Press, June 4, 2010, DOI 10.1074/mcp.R000007-MCP201

Given the complementary nature of these approaches, it would seem natural for there to be *in silico* tools that support the interactive exploration of structural biology within the context of the proteome, and of the results of proteomics experiments from a structural perspective. However, although several studies that link proteomics to structure have been published (1–5), there are few existing tools for the interactive, integrated exploration of these complementary types of data. The structural biology and proteomics communities each have a set of commonly used interactive visualization programs, and it would be useful to investigate how these tools could work together and how they could be more tightly linked.

INTERACTIVE TOOLS FOR PROTEIN SYSTEMS DATA

Network visualization and analysis tools are commonly used to interact with proteomics data. This makes good sense: proteomics data are often associated with pathways or protein interactions, and both of these are easily visualized as networks. Even types of data not normally viewed as networks (e.g. microarray results) are often painted onto signaling, metabolic, or other pathways or protein interaction networks for visualization and analysis.

A number of visualization and analysis tools are used for protein networks. The most commonly used tool is certainly Cytoscape (6, 7), but others such as VisANT (8), Osprey (9), BioLayout Express3D (10), Arena3D (11), and PATIKA (12) are also cited. In the commercial space, Pathway Studio is commonly used (13). For a useful review of the various biological network analysis tools, see Pavlopoulos *et al.* (14).

INTERACTIVE TOOLS FOR STRUCTURAL BIOLOGY

Structural visualization and analysis have a very long and rich history, and a discussion of the various molecular visualization and analysis packages is beyond the scope of this article. The most common stand-alone molecular visualization packages are PyMOL (15), VMD (16), and UCSF Chimera (17). Jmol (18) and Rasmol (19) are often used as web add-ons for structural visualization, and the Research Collaboratory for Structural Bioinformatics and the National Center for Biotechnology Information both have their own viewers (Protein Workshop (20) and Cn3D (21), respectively). In the commercial space, Sybyl from Tripos and Discovery Studio from Accelrys are widely used.

INTERACTIVE TOOLS LINKING STRUCTURAL AND PROTEOMICS DATA

To date, there are very few tools that provide any kind of interactive linkage between proteomics data sets and the

structures of proteins. STRING (22) is a web service that provides the user with a protein-protein interaction network. The user may click on a node to reveal more information about the protein, including a static image of the three-dimensional structure if known. Clicking the image takes the user to the European Molecular Biology Laboratory-European Bioinformatics Institute web entry for that structure, which allows interactive visualization with Jmol.

A different approach is taken by structureViz (23), a plug-in to Cytoscape that loads the structures for network nodes designated by the user into UCSF Chimera for interactive three-dimensional visualization and analysis. Interaction is bi-directional so that selecting a structure in Chimera will select the appropriate node in Cytoscape.

CYTOSCAPE AND CHIMERA

As discussed above, a variety of computational tools are available for research in proteomics and structural biology, and it is beyond the scope of this article to provide a detailed comparison between them. Some of these tools may be used together to “drill down” from the proteomics, network-oriented view to a structural view. One pair of tools that may be used together in this manner is Cytoscape (6, 7), an open source package widely used to visualize and analyze networks, and Chimera (17), a well-supported and widely distributed academic molecular visualization and analysis package. We explore how Cytoscape and Chimera might be used together by presenting three example research scenarios. Our focus is on the computational tools rather than on the specific data; the scenarios are based on previously published studies, and the results are not meant to represent novel findings. It is also the case that both Chimera and Cytoscape are relatively sophisticated tools with many features that may require some effort to fully master. Our intent is not to illustrate all of the features available in these tools but rather to provide examples of how they can be applied to gain insight into scientific problems. Lastly, it is difficult to convey the interactive nature of these tools using static images. To give a basic idea of the dynamic nature of the systems under study, we have provided two animations as supplemental Movies 1 and 2.

SCENARIOS

The first two scenarios deal with glioblastoma multiforme (GBM),¹ the most common and aggressive brain tumor in

humans (24, 25). Glioblastoma multiforme was also the first of the cancer types to undergo comprehensive genomic characterization by The Cancer Genome Atlas (TCGA) project (26). The TCGA glioblastoma data set consists of three types of data: copy number variation of 17,789 genes for 206 glioblastoma cases, mRNA expression for the same 17,789 genes and 206 cases, and mutation data for 601 sequenced genes for 91 of the cases. In scenario 1, we use Cytoscape to explore a curated signaling pathway obtained from the TCGA data portal. From the GBM mutation data mapped onto the pathway, we choose one mutation of interest and drill down into Chimera to view the possible structural implications.

Scenario 2 focuses on isocitrate dehydrogenase 1 (IDH1), a metabolic enzyme that has been found mutated in glioblastoma (27). Using networks, we explore the function of wild-type and mutant IDH1 to hypothesize how the mutation might relate to glioblastoma.

Finally, in scenario 3, we look at a protein-protein interaction data set from the budding yeast *Saccharomyces cerevisiae* and examine how these data can support the modeling of large protein complexes (see the articles by Lasker *et al.* (75) and Förster *et al.* (76) in this issue). This data set has been annotated with available structural data, and we show how this information might be used in conjunction with automated fitting within Chimera.

Scenario 1: Mutations in Signaling Proteins Associated with Glioblastoma Multiforme—There are a number of repositories of curated pathways that can be loaded into Cytoscape, including the Kyoto Encyclopedia of Genes and Genomes (28–30), Reactome (31), Pathway Commons, BioCyc (32), the NCI-Nature Pathway Interaction Database (33), and WikiPathways (34, 35). In addition, there are many repositories of protein-protein interaction data sets such as the Human Protein Reference Database (36), Pathway Commons, and STRING (22) that may be used to augment existing pathways with additional interaction partners. This scenario uses a curated signaling pathway provided on the TCGA data portal. This pathway represents the most frequently altered genes in glioblastoma based on the TCGA phase I data (26). In addition to the curated pathway, the TCGA data portal provides downloads of the three data sets that can be used to annotate the pathway: expression, copy number variation, and mutations. Supplemental Fig. 1 shows a screenshot of Cytoscape with the TCGA-curated pathway for glioblastoma loaded and provides a description of the user interface for Cytoscape. A Cytoscape session file with the TCGA pathway is included as supplemental Data 1.

First we explore the expression profile of each of the genes across all of the tumor patients. The differential regulation of gene expression has been associated with a large number of diseases (37, 38) and can implicate specific genes. The usual

¹ The abbreviations used are: GBM, glioblastoma multiforme; TCGA, The Cancer Genome Atlas; PDB, Protein Data Bank; IDH, isocitrate dehydrogenase; CDKN2A, cyclin-dependent kinase inhibitor 2A; CDK4, cyclin-dependent kinase 4; PDGFRA, platelet-derived growth factor receptor, α polypeptide; ERBB3, v-erb-b2 erythroblastic leukemia viral oncogene homolog 3, also known as HER3; TP53, tumor protein 53; PTEN, phosphatase and tensin homolog; EGFR, epidermal growth factor receptor; NF1, neurofibromin 1; FGFR1, basic fibroblast growth factor receptor 1; HIF-1 α , hypoxia-inducible

factor 1, α subunit; 2HG, 2-hydroxyglutarate; GO, gene ontology; RPB1–12, RNA polymerase subunits 1–12; SEA, Similarity Ensemble Approach; PE, purification enrichment; MCL, Markov cluster.

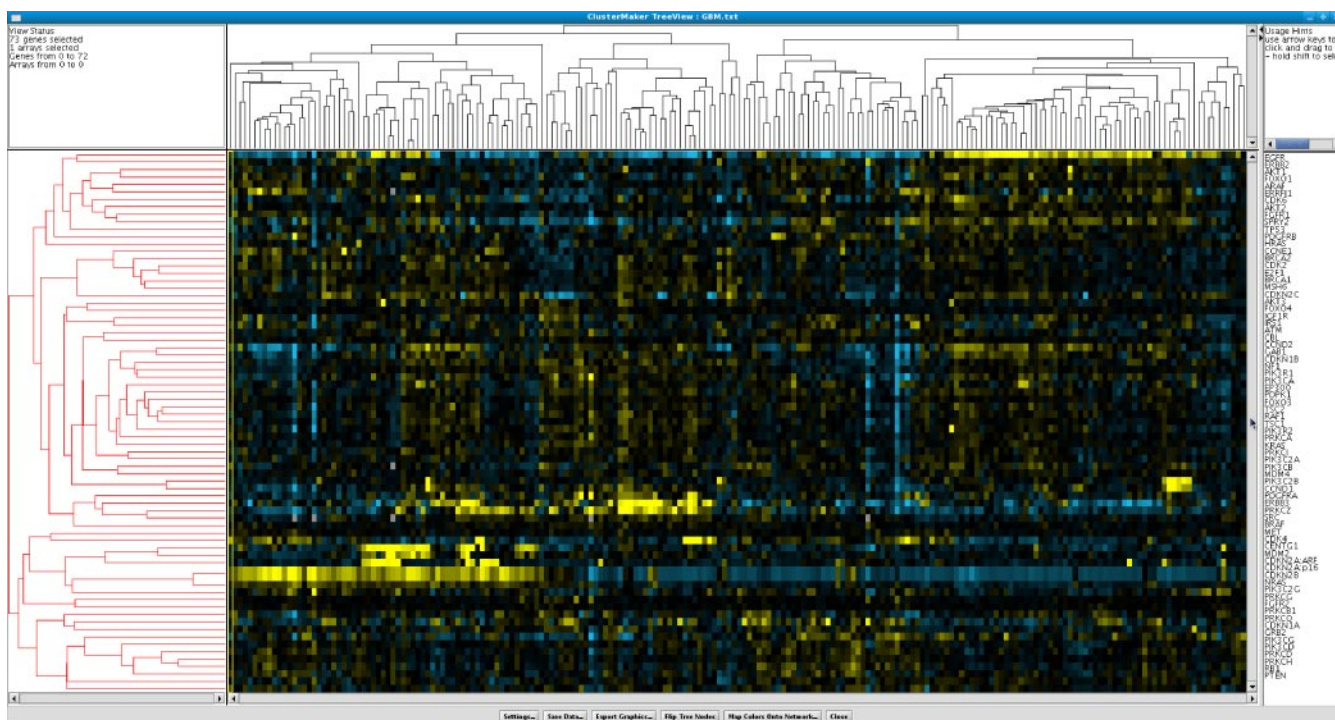


FIG. 1. **TCGA mRNA expression data.** This screenshot of the clusterMaker hierarchical cluster results shows the mRNA expression data across all tumors in the TCGA study and all genes in the TCGA GBM pathway.

mechanism to view differential gene expression across multiple genes and conditions is to hierarchically cluster the data and view the results as a heat map with dendrograms representing the clusters for both genes and conditions (39), where each tumor represents a different condition in this example. After annotating the TCGA pathway with the mRNA expression results, we can use the Cytoscape clusterMaker plug-in to perform the clustering (Fig. 1). As described in the TCGA 2008 report, this clustering does not lead to any obvious conclusions; that is, none of the genes are overexpressed or underexpressed in all (or even most) of the tumors.

On the other hand, looking at the clustering of tumors, we can see two broad categories: those overexpressing *CDKN2A* or *CDK4* and those underexpressing *PDGFRA/ERBB3* and *CDKN2A*. Although these groups are discernable, there remain certain inconsistencies within the groups that prevent a clear categorization of the tumors.

To view differential mRNA expression in the context of the pathway, we can animate the coloring of the nodes in the pathway using the “Map colors to network” capability of the clusterMaker plug-in (supplemental Movie 1), and it can be seen that for each tumor some sets of genes are either over- or underexpressed, but there is no readily discernable pattern. The lack of expression patterns might lead to an exploration of copy number variation or mutations. Like the mRNA expression data, copy number variations can be analyzed by clustering (Fig. 2), although a more detailed analysis, including structural data where available, may be required for the individual mutations.

For the mutation analysis, we used the same TCGA pathway, annotated it with the known structures for each gene product from the Protein Data Bank (PDB) (40, 41), and imported the mutation data from the TCGA data portal. We added to the imported data an additional column to represent the percentage of tumors that were mutated for each gene. Fig. 3 is an export from Cytoscape showing each gene colored by the percentage of sequenced tumors showing mutations for that gene. Among the most strongly colored genes are *TP53*, *PTEN*, *EGFR*, and *NF1*, which have been identified previously as mutated in many tumors. Interestingly, the most frequently mutated gene, *TP53*, is only mutated in 34% of the tumors, and the tyrosine phosphatase *PTEN* is only modified in 30.7% of the tumors. One of the less frequently mutated proteins, basic fibroblast growth factor receptor 1 (FGFR1), has been well studied, and partial structures of the protein are available in the PDB. Although this protein is less frequently mutated than some of the other genes, the nature of the mutation and availability of structures provide an interesting example for our structural analysis.

FGFR1 is a receptor tyrosine kinase. Like other receptor tyrosine kinases, it comprises an extracellular ligand-binding part, a single transmembrane-spanning segment, and an intracellular part that includes a protein-tyrosine kinase domain. Ligands such as fibroblast growth factor that activate FGFR1 cause receptor dimerization and autophosphorylation across the dimer interface. Autophosphorylation shifts the kinase domain into an active state. The activated recep-

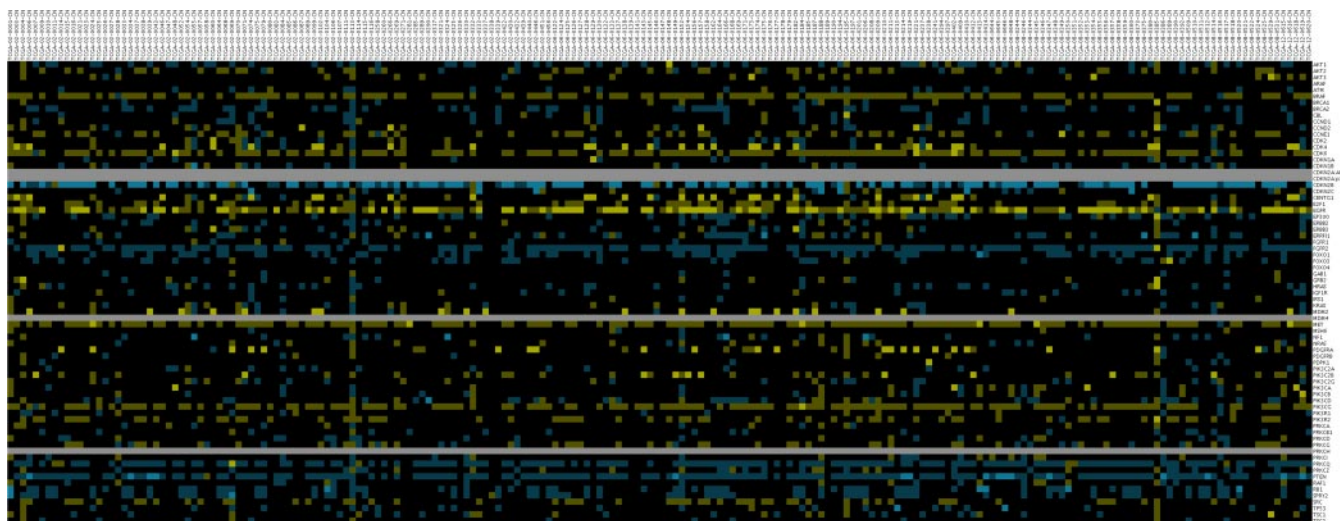


FIG. 2. **TCGA copy number variation data.** This heat map shows the copy number variations for genes in the TCGA glioblastoma pathway across all of the tumors. The colors indicate the copy number variation noted: *bright blue*, homozygous deletion; *blue*, hemizygous deletion; *black*, no variation; *gray*, not measured; *yellow*, gain; *bright yellow*, high level amplification.

tor goes on to bind and phosphorylate several downstream partners (42).

FGFR1 signaling is involved in growth and proliferation, and overactivity has been associated with various cancers. Other glioblastoma-associated mutations in FGFR1 have been discussed previously (43, 44), so here we only address the mutation K656E found in the TCGA study. The structureViz Cytoscape plug-in can be used to load structures into Chimera for analysis. The two structures of interest are the structure of the FGFR1 kinase domain in the inactive state (for example, PDB code 3c4f (45)) and the phosphorylated, activated state (for example, PDB code 3gqi (46)) (Fig. 4).

Multiple tyrosine residues in FGFR1 are autophosphorylated, and the phosphorylations occur in a stereotyped order with successive stages incrementally increasing kinase activity (44). Remarkably, given their adjacency in sequence, Tyr-653 is the first to be phosphorylated, but Tyr-654 is phosphorylated last, after several other tyrosines (44). Tyr-653, Tyr-654, and the glioblastoma mutation site Lys-656 are all within the “activation loop,” which undergoes a large conformational change.

The two structures of the FGFR1 kinase domain can be superimposed in Chimera to show the magnitude of the conformational change (Fig. 4). In the activated conformation, Lys-656 is hydrogen-bonded (H-bonded) to phospho-Tyr-654 as indicated by the *red dashed line* in Fig. 5. Mutation of this lysine to glutamate, which is negatively charged, could mimic or interfere with phosphorylation, or otherwise affect conformation or molecular recognition. Of note, mutation from a positively charged lysine to a negatively charged glutamate gives a net change in charge of -2 , about the same (depending on pH) as a single tyrosine phosphorylation. Mimicry of phospho-Tyr-654 is compelling because Glu-656 and unmodified Tyr-654 might still form an H-bond and occupy a similar volume and thus increase kinase ac-

tivity in the absence of the final autophosphorylation. The negative potential from Glu-656 might also enhance earlier stages of activation.

Several types of analyses can be performed in Chimera. As alluded to above, the structures can be superimposed. In this case, the MatchMaker tool was used; it automatically pairs residues based on an initial sequence alignment and iterates the fit so that only the well-matched portions are used in the final superposition (47). Distances between activation loop residues in the two states can be evaluated as a measure of the conformational change. The change can be represented visually by not only static images of superimposed structures but as an animation generated by “morphing” between the different conformations. Chimera analyses suitable for the individual structures include finding H-bonds and contacts of residues (thus discerning their structural and functional roles and, by extension, what might happen if they were mutated) and coloring the molecular surface by electrostatic potential. Another way to assess the importance of a residue is by calculating its conservation in a sequence alignment using the Multalign Viewer tool of Chimera (47); several measures of conservation can be computed, and the alignment can be read in from an external file or generated within Chimera, for example by performing a BLAST (Basic Local Alignment Search Tool) search. Finally, a first-order, crude analysis of a mutation can be performed by “swapping” one residue type for another in the structure with the Rotamers tool. All of these analyses can be performed interactively and take at most a few seconds, and so it is quite reasonable to try out “what if” ideas in an exploratory fashion as a means of potentially generating new hypotheses.

Scenario 2: Mutations in IDH1 Associated with Glioblastoma Multiforme—This scenario begins from the structural perspective. The enzyme IDH1 has been found commonly

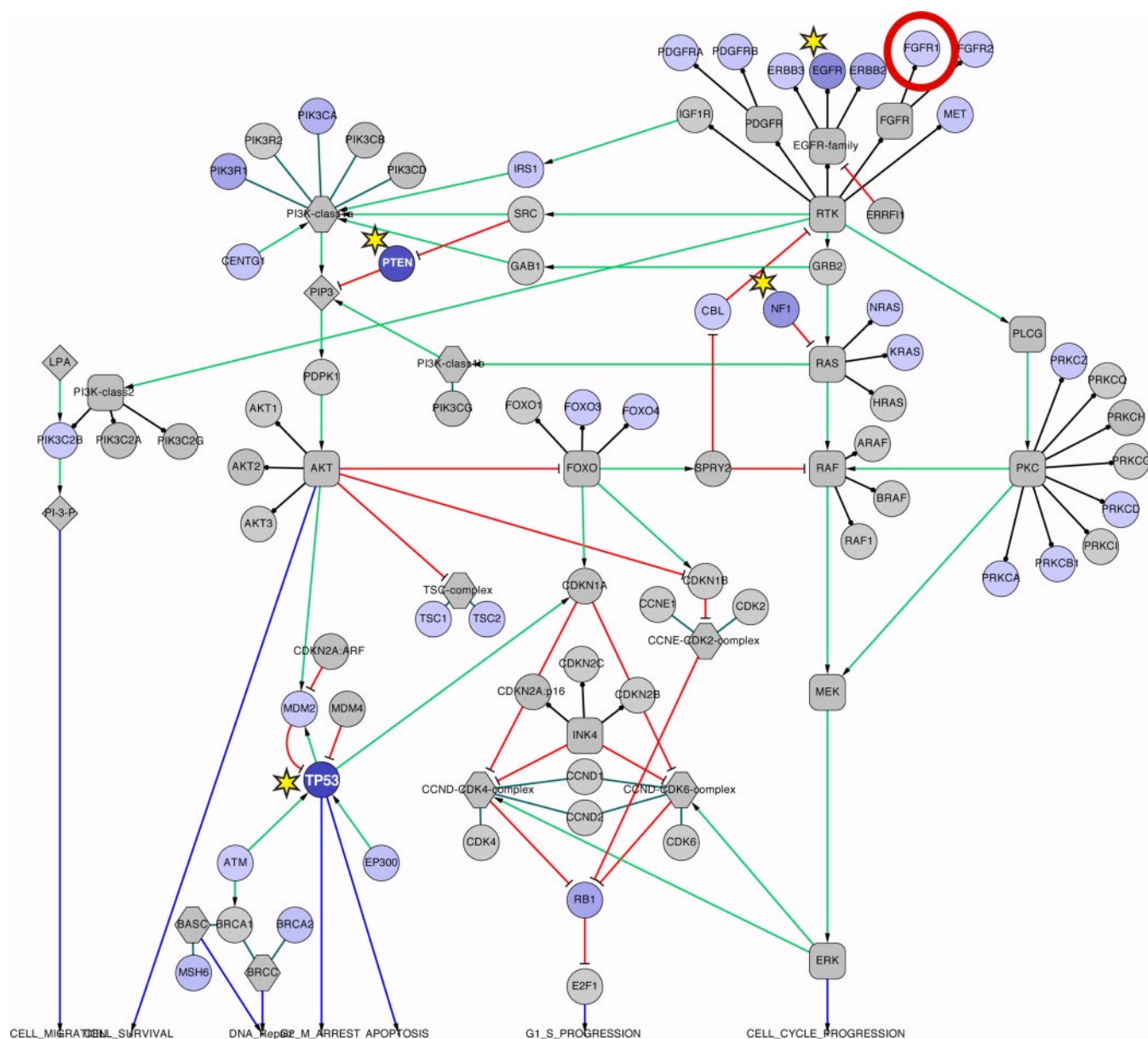


FIG. 3. **TCGA glioblastoma pathway.** This Cytoscape export of the TCGA glioblastoma pathway shows the protein nodes colored by the percentage of the 91 sequenced tumors that showed mutations for that protein. Protein nodes with *yellow stars* indicate the most highly mutated nodes. The node in the *red circle* (FGFR1) is the protein studied for scenario 1.

mutated at Arg-132 in human glioblastoma multiforme (27). IDH1 converts isocitrate to α -ketoglutarate by oxidative decarboxylation. Viewing the structure of IDH1 (PDB code 1t0l) (48) reveals that the mutation is in the active site of the enzyme (Fig. 5); Arg-132 forms hydrogen bonds (*red dashed lines*) with the substrate, isocitrate. Assays of the R132H, R132C, and R132S mutants of IDH1 *in vitro* found substantial reductions in both affinity for substrate and catalysis of the normal reaction, and expression in cultured cells showed reduced levels of the product α -ketoglutarate (49). The lower levels of α -ketoglutarate led to higher levels of the hypoxia-inducible factor 1, α subunit (HIF-1 α). This might explain the

cancer association, as HIF-1 α facilitates tumor growth under low-oxygen conditions.

However, Dang *et al.* (50) argued that a gain of function by IDH1 mutants rather than a loss is important for cancer causation and that the effect is mediated by an alternative reaction product or “oncometabolite,” 2-hydroxyglutarate (2HG). Several lines of evidence were presented. (a) Metabolite profiling identified 2HG as significantly higher in cells expressing R132H *versus* wild-type IDH1. (b) The R132H, R132C, R132L, and R132S mutants of IDH1 catalyzed the reduction of α -ketoglutarate to 2HG *in vitro* (*in vivo*, the remaining wild-type copy of IDH1 could supply the α -ketoglutarate). (c) α -Ketoglutarate

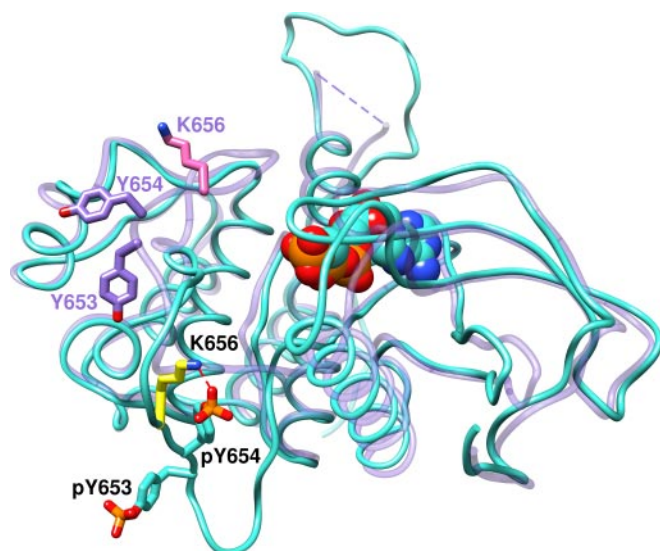


FIG. 4. Superimposed structures of FGFR1 kinase domain. The backbones are shown as “licorice” ribbons, and the side chains of interest are shown as sticks. *pY*, phosphotyrosine. Heteroatoms are colored by element: oxygen, red; nitrogen, blue; phosphorus, orange. The activated structure (PDB code 3gqi, chain A) is shown in turquoise with black labels; in this conformation, the Lys-656 side chain (yellow carbons) is H-bonded to phospho-Tyr-654 as indicated by the red dashed line. The activated structure also includes an ATP analog, displayed here in the space-filling representation. The inactive conformation (PDB code 3c4f, chain A) is shown with a transparent purple backbone and purple labels. (The purple dashed line indicates structure information missing from the PDB file.) In the inactive conformation, the Lys-656 side chain (pink carbons) is not forming any intramolecular H-bonds. The Chimera command file for setting up this image is available as supplemental Data 2. The commands in the file set up everything except the orientation and labels, which although scriptable are normally generated interactively.

utamate levels were not statistically different between tumors with and without *IDH1* mutations, whereas 2HG levels were higher in the *IDH1* mutant tumors. (d) Patients with elevated 2HG caused by mutations in a different enzyme also have an increased risk of developing brain tumors. How 2HG itself increases cancer risk is not known, but the authors listed some possible mechanisms (50).

The experimentally determined structures of native *IDH1* (PDB code 1t01) and the R132H mutant (PDB code 3inm) can be aligned in Chimera using the MatchMaker extension. The structure of the R132H mutant is highly similar overall to that of the wild type, with the most obvious change being the loss of the arginine side chain at position 132 and its interactions (Fig. 5). The main image in Fig. 5 shows the wild-type structure, whereas the inset shows the mutant (PDB code 3inm) with isocitrate modeled into the structure based on the MatchMaker superposition with the wild-type structure. In this modeled complex, the His-132 side chain is too far from the isocitrate to H-bond with it, as assessed by the Chimera FindHBond tool with default settings.

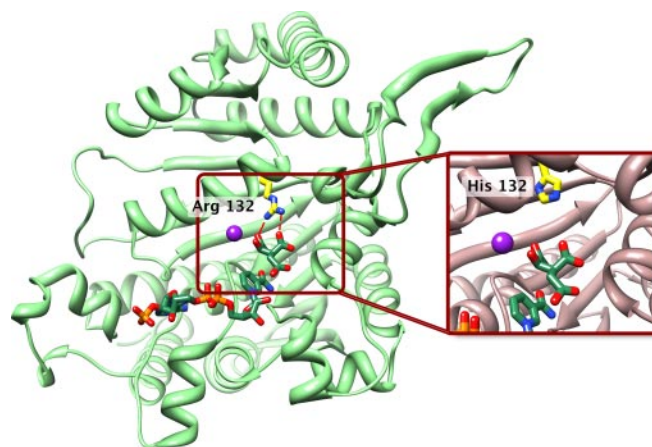


FIG. 5. Wild-type and mutant *IDH1* structures. The wild-type structure (PDB code 1t01, chain D) is shown on the left, with the substrate isocitrate, cofactor NADP (both dark green with heteroatom color coding as in Fig. 5), and a calcium ion (purple) bound in the active site. Dashed red lines show H-bonds between Arg-132 and isocitrate identified with the FindHBond extension in Chimera. The inset at right shows the mutated protein (PDB code 3inm, chain A) in which residue 132 is a histidine, with isocitrate and the other bound ligands modeled into the structure based on the MatchMaker superposition with the wild-type structure. In this modeled complex, the His-132 side chain is too far from the isocitrate to H-bond with it, as assessed by FindHBond.

Several metabolic pathways involve the isocitrate dehydrogenases (besides *IDH1*, there are *IDH2* and *IDH3* isoforms), including the tricarboxylic acid cycle (also known as the Krebs cycle or citric acid cycle), the main energy-producing cycle in the cell. *IDH1* is not found in the mitochondria in humans and hence is not directly involved in the tricarboxylic acid cycle. *IDH1* resides in the cytosol and peroxisome and is NADP⁽⁺⁾-dependent. It is associated with NADPH regeneration and glutathione metabolism. We could utilize Cytoscape to place *IDH1* in the appropriate pathway, but our main concern is not so much the enzyme but the oncometabolite 2HG that the mutant enzyme has been shown to produce (50). We would like to identify proteins that bind this compound to reveal (a) potential explanations of why it increases the risk of cancer or (b) entities in which mutations might lead to increased levels of 2HG and thus increased risk. The Similarity Ensemble Approach (SEA) provides a resource to search for proteins based on similarity of the compounds they bind (51). We performed a search on the SEA web site for “C(CC(=O)O)C(C(=O)O)O,” the SMILES (52) representation for 2-hydroxyglutarate. Working with the author of SEA (53), we loaded a version of the SEA network, “chembl,” into Cytoscape and selected a subset of the proteins that were returned as significant by our search. This gave us a small network for viewing some of the proteins known to bind to small molecules similar to 2HG (Fig. 6A). The chemViz plug-in to Cytoscape allows inspecting the chemical properties of the ligands of proteins selected from the SEA network. Properties of the small molecule ligands include two-dimensional structure, numbers of H-bond acceptors and donors, molecular

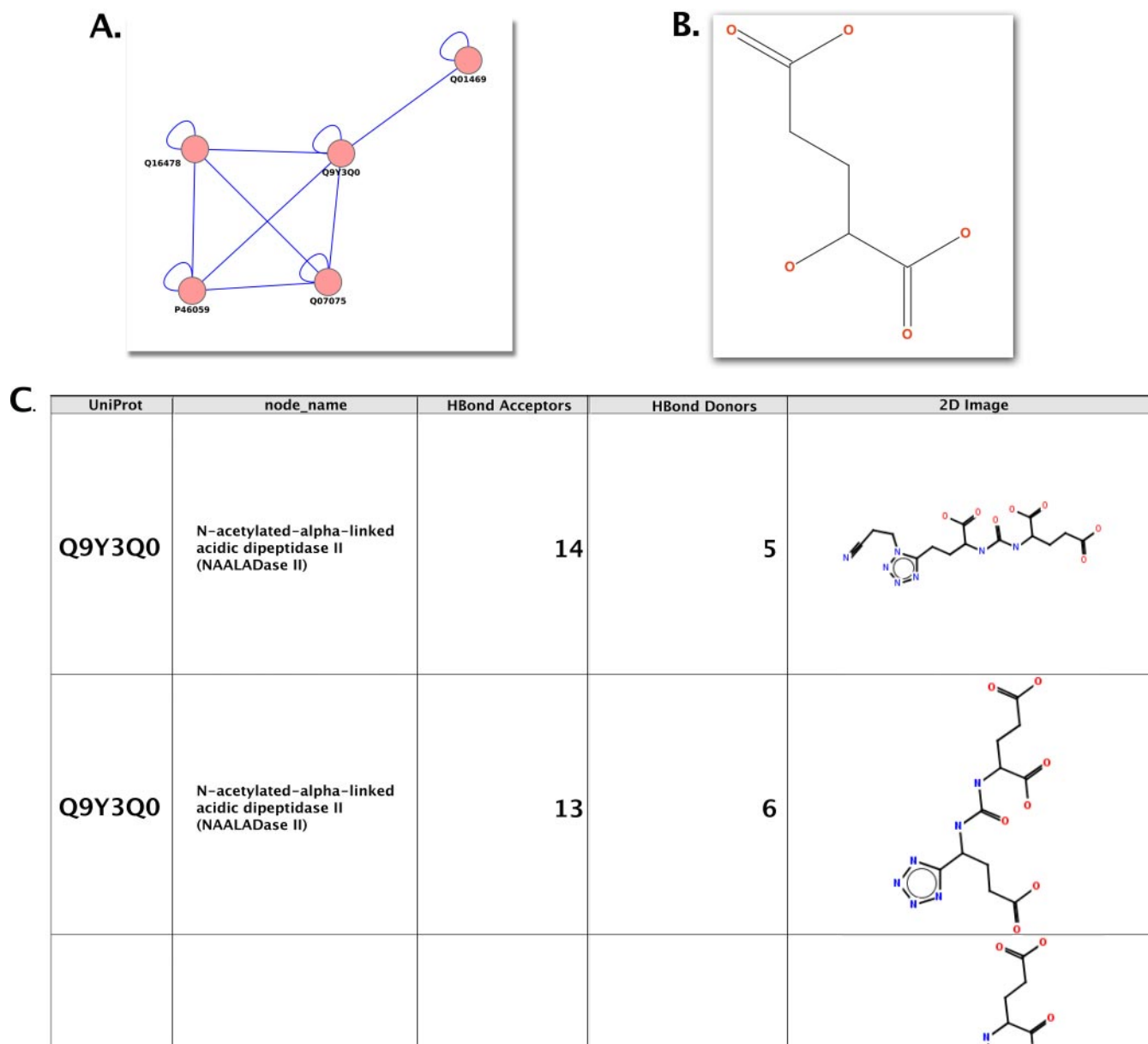


FIG. 6. **2HG similarity.** A portion of the chembl network (A) shows significant hits from a SEA search for 2-hydroxyglutarate (B). Nodes are proteins with known ligands that match the structure of 2-hydroxyglutarate, and the edges represent the similarity between the ligands known to bind to the proteins. The table (C) was generated with the chemViz Cytoscape plug-in and shows a portion of the compounds known to target the protein with UniProt accession number Q9Y3Q0 (the text has been expanded for clarity).

weight, and several other widely used cheminformatics descriptors (Fig. 6C).

The five proteins in our network do not hold many surprises. Q9Y3Q0 is a membrane carboxypeptidase that releases an unsubstituted C-terminal glutamyl residue (54). Q16478 is a glutamate-gated ion channel found in the cell junction and postsynaptic membrane. Q01469 is a fatty acid-binding protein, and P46059 is a peptide transporter involved in digestion (55) and protein transport.

A more interesting result is the glutamyl aminopeptidase Q07075, also known as aminopeptidase A. This protein is

known to regulate angiotensins, including those in the brain renin-angiotensin system (56), and has been implicated for its role in brain tumor-associated vessels (57). Earlier studies also suggest that aminopeptidase A has a role in differentiation and cell proliferation at least in renal tissues (58). This may be an interesting target for 2HG both in terms of its potential role in differentiation and its known activity in regulating angiotensin. We have not yet followed up on this finding, as our immediate goal here was to illustrate how tools can be used to explore and discover possibly interesting proteins not previously associated with glioblastoma multiforme.

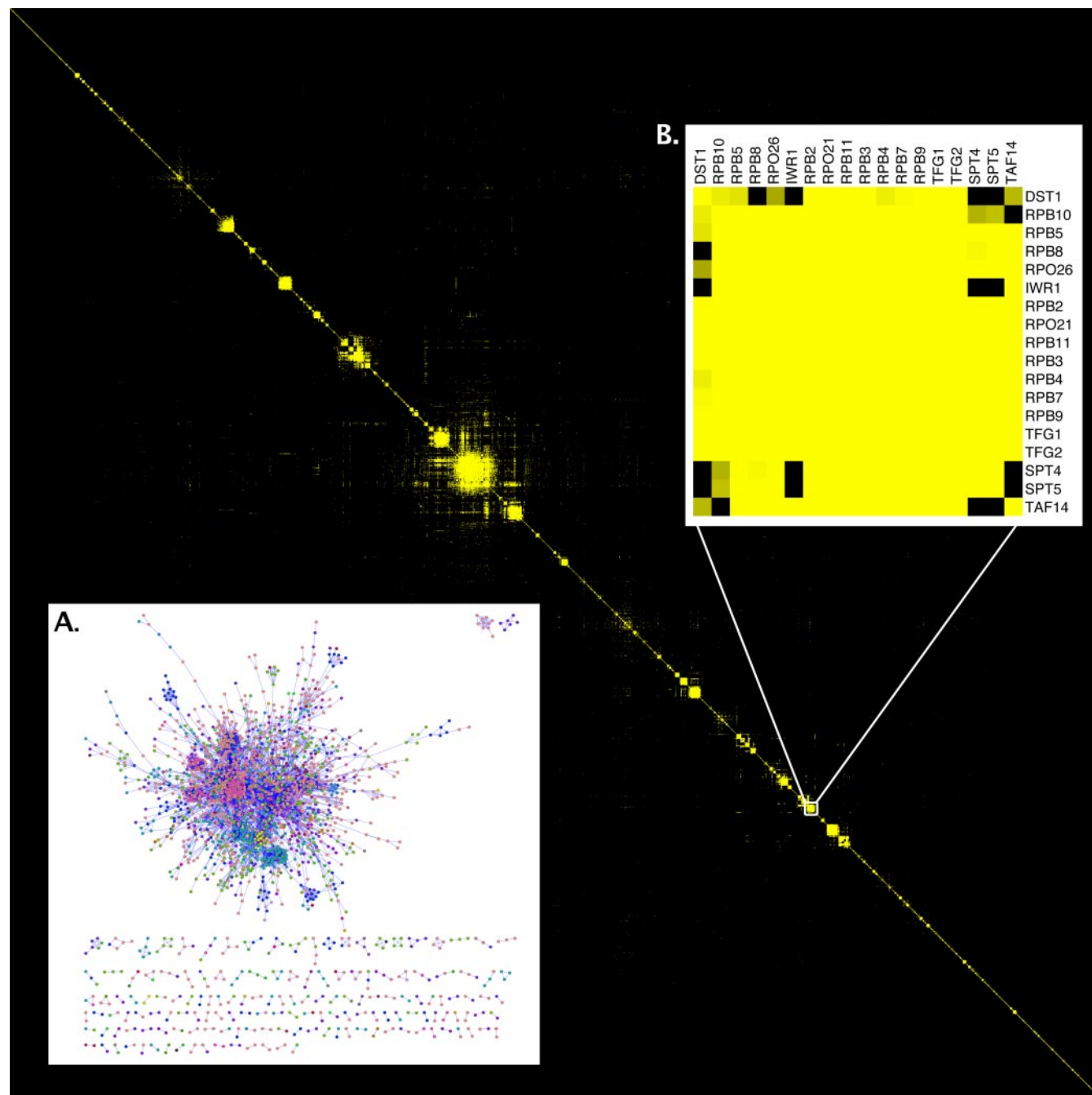
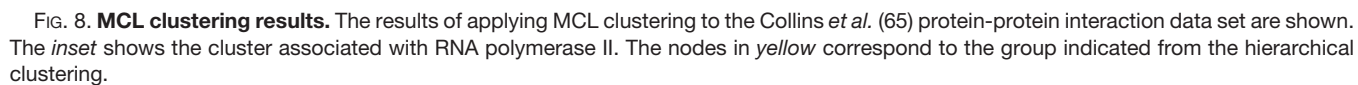


FIG. 7. **Collins et al. (65) protein-protein interaction data.** A hierarchical clustering of the Collins et al. (65) protein-protein interaction network for *S. cerevisiae* is shown. *Inset A*, in the lower left is the non-clustered protein-protein interaction network with the nodes colored by the GO annotation for biological function. *Inset B* shows a close-up of the cluster for RNA polymerase II.

Scenario 3: Using Proteomics to Inform Structural Modeling—In our final scenario, we explore the use of proteomics data as an input to modeling structures of protein complexes. The computational techniques for fitting atomic structures into density maps and the role of protein-protein interaction data are described elsewhere in this issue (Lasker et al. (75) and Förster et al. (76) in this issue). In this section, we examine how interactive tools might be used prior to the final auto-

mated fitting by a program such as IMP (59, 60) or Segger (61).

To begin our scenario, we use a protein-protein interaction data set for the budding yeast *S. cerevisiae* (downloaded on December 18, 2008) that contains the combined data from three previous studies by Ho et al. (62), Gavin et al. (63), and Krogan et al. (64). Each of these studies used tandem affinity purification followed by mass spectrometry analysis to create



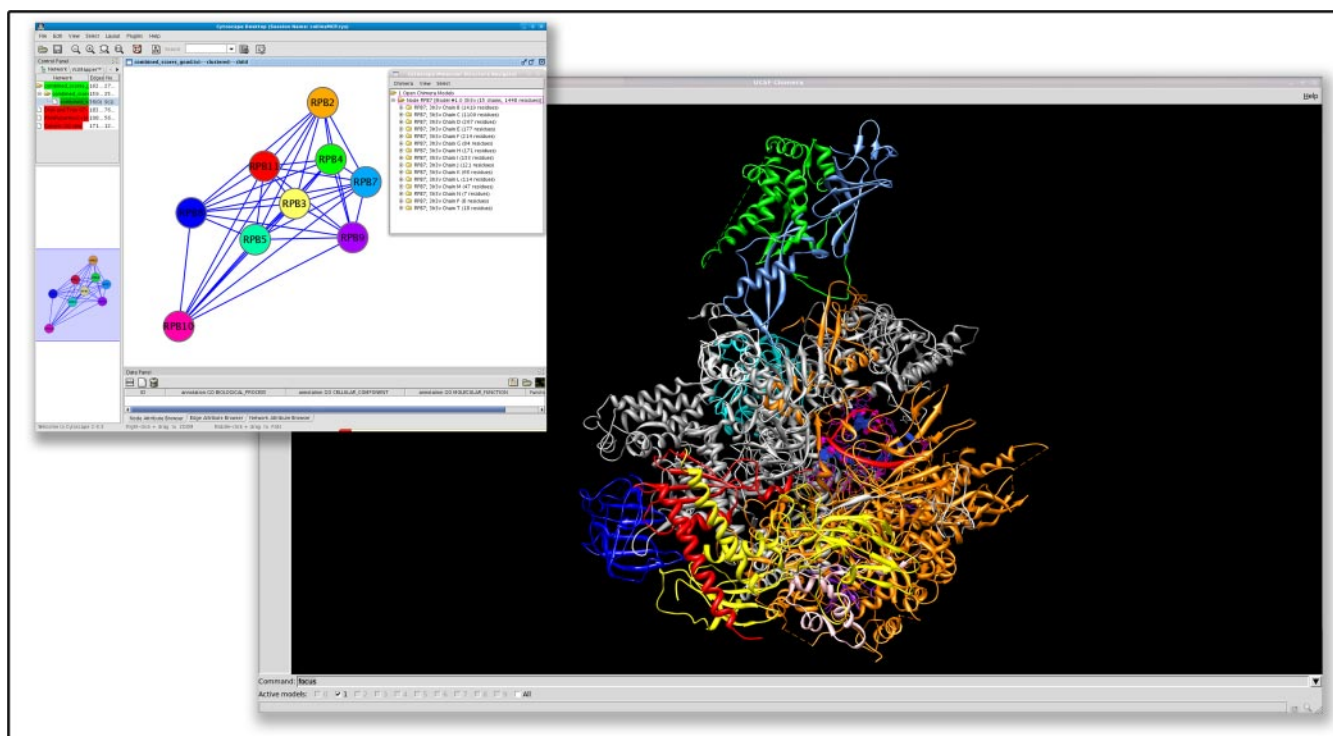


FIG. 9. **Cytoscape and Chimera views of RNA polymerase II.** The nine subunits of RNA polymerase II in the Collins *et al.* (65) protein-protein interaction data set are shown with the node colors randomly assigned. The structure has been opened with the structureViz Cytoscape plug-in (23), and the colors of the individual chains have been colored to match the node colors. The colors of RPB1, RPB6, and RPB12 have been set to *dark gray*, *gray*, and *white*, respectively.

networks of interactions between the proteins in the study. Collins *et al.* (65) combined these data sets to increase coverage and improve accuracy. Gene ontology (GO) (66) annotations were added using the Cytoscape “Import ontology and annotations” capability. The resulting network is shown in Fig. 7A, which has nodes (proteins) colored by the GO biological process and positioned using the Cytoscape force-directed layout. A Cytoscape session file with these data is included as supplemental Data 3.

Although the network is useful, it is difficult to see the individual relationships between the proteins in a complex or to distinguish strong interactions from weaker interactions. The network can be clustered hierarchically with clusterMaker by selecting the PE scores as the source for the array data (Fig. 7). An alternative approach is to use the implementation of Markov cluster (MCL) in clusterMaker (67) to break the network into distinct groups (Fig. 8). Either type of clustering yields groups of nodes where a single group represents a set of proteins that have been found to interact and, in most cases, form a complex (64). clusterMaker links different displays of the same data so that a selection in the hierarchical clustering results will be reflected in the network view. This allows the user to interactively interrogate both sets of results.

RNA polymerase II is used to illustrate how protein-protein interaction data can feed into the structural modeling of pro-

tein complexes (Fig. 7 and Fig. 8, *inset*). Using both MCL and hierarchical clustering, we can quickly isolate the proteins involved in the RNA polymerase II complex. As with the previous networks, this network has also been annotated with the available PDB structures. All of the available structures include multiple chains, reflecting one or more of the components of the complex. To provide a more useful connection to three-dimensional structure, the nodes have been further annotated with the appropriate chain. With structureViz, the individual chains (individual components of the complex) can be assigned colors in Chimera (Fig. 9).

First, we loaded a structure of the yeast RNA polymerase II complex that includes all 12 subunits and a poly(A) signal in the active site (PDB code 3h3v) (68). Fig. 9 shows the structure in Chimera along with a network in Cytoscape for nine of the 12 protein subunits in RNA polymerase II. The nodes in the network have been colored to match the colors of the chains in the complex. The relative position of the nodes in the network reflects the strength of the interaction (PE score) (65). As might be expected, the PE score between RPB4 and RPB7 is very high (*green* and *medium blue* nodes near the *right* side of the network), and this is reflected in the structure (*green* and *medium blue* chains at the *top* of the structure). These two subunits bind very closely together to form a heterodimer that can be dissociated from the rest of the complex (69, 70). Together they make up the “stalk” of RNA polymerase II.

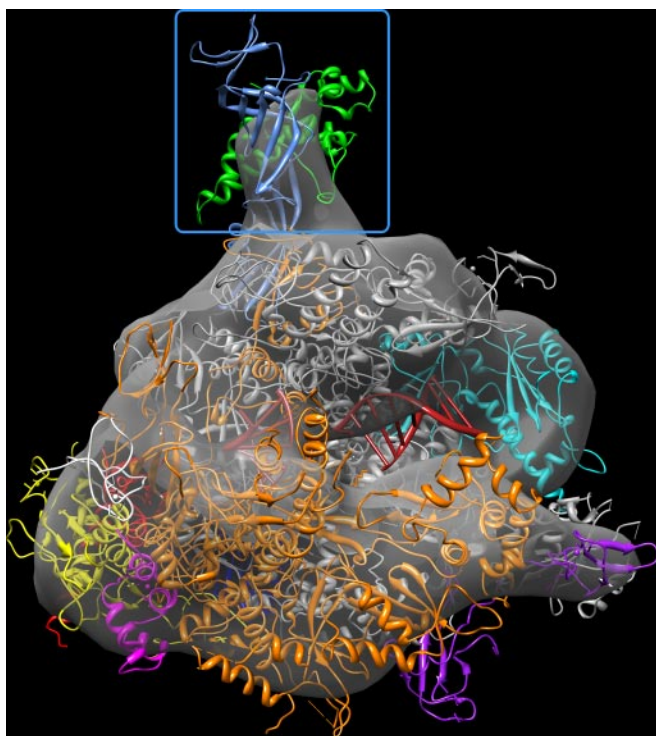


FIG. 10. Yeast RNA polymerase II with human RPB4-RPB7 fit. The yeast RNA polymerase II structure fit into a cryoelectron microscopy map of the human complex is shown. The stalk structures (in the blue square) are human RPB4-RPB7 that have been placed by aligning them with the homologous yeast structures using the MatchMaker tool in Chimera. See the Chimera tutorial on density maps (online at the Chimera web site) for instructions on how to duplicate this figure.

To see how this structure compares with the human RNA polymerase II complex, we can use the volume viewing capabilities of Chimera (71) to load a cryoelectron microscopy map of the full human RNA polymerase II complex (Electron Microscopy Data Bank code 1283) (72) from the Electron Microscopy Data Bank (73). This map is part of a series of three showing the conformational flexibility of the human complex.

The next step is to manually position the structure into the map and then use the Fit in Map tool in Chimera to perform a local optimization. The results can be refined further by loading the human RPB4-RPB7 complex (PDB code 2c35) (74) and superimposing those chains with the corresponding yeast subunits using MatchMaker (described above). The yeast stalk can then be hidden, leaving only the human stalk modeled together with the rest of the yeast complex (Fig. 10). Lasker *et al.* (75) in this issue used modeled structures for the human complex subunits and computationally fit them into the same map using the IMP multifit module (59).

DISCUSSION

The value of combining proteomics and structural biology to understand biological processes is unquestioned. Any with remaining doubts should be convinced by the articles in this

special issue. We hope we have demonstrated the value of *in silico* tools that interactively link these data sets to allow researchers to explore them in context and to use both types of data synergistically for hypothesis generation. The tools presented in the three scenarios are only partially linked, and we hope that increased use will allow us to understand how structural biologists will want to “zoom out” to view their structures in a pathway or interaction context and how the proteomics community will want to “zoom in” to analyze structural data.

In the near future, there are two areas in which we hope to significantly improve the combined use of these tools. First, it is currently very difficult to annotate a Cytoscape network with PDB (structure database) identifiers, which is a prerequisite for using structureViz. We hope to provide a Cytoscape plug-in (or extend structureViz) to search the PDB for each protein in a network and add the list of PDB identifiers to the node attributes.

Second, there are several useful web-based resources that are cumbersome to access from within Cytoscape. Many of these resources support a web service that allows searching for and downloading annotation data and, in many cases, downloading the results of analyses performed on behalf of the user. Integrating some of these services (for example, SEA) into Cytoscape would allow easier access.

Finally, we would like to enable structural biologists using Chimera to search for relevant pathways or networks and load them into Cytoscape. Both views should be interactive, and the structures in Chimera and corresponding nodes in Cytoscape should be automatically linked so that selections in one program will propagate to the other.

We encourage researchers who focus mainly on structural biology or mainly on proteomics to “cross the aisle” where possible and, by identifying deficiencies in existing computational tools or unmet needs, to drive the development of new and improved analysis and visualization tools. New types of data may bring new insights, and the synthesis of proteomics and structural biology will certainly yield more than the sum of its parts.

* This work was supported, in whole or in part, by National Institutes of Health Grant P41 RR01081 from the National Center for Research Resources.

§ This article contains supplemental Fig. 1, Movies 1 and 2, and Data 1–3.

‡ To whom correspondence should be addressed: Resource for Biocomputing, Visualization, and Informatics, Dept. of Pharmaceutical Chemistry, University of California, 600 16th St., M/S 2240, San Francisco, CA 94158-2517. Tel.: 415-476-2299; Fax: 415-502-1755; E-mail: tef@cgl.ucsf.edu.

REFERENCES

1. Kühner, S., van Noort, V., Betts, M. J., Leo-Macias, A., Batisse, C., Rode, M., Yamada, T., Maier, T., Bader, S., Beltran-Alvarez, P., Castaño-Diez, D., Chen, W. H., Devos, D., Güell, M., Norambuena, T., Racke, I., Rybin, V., Schmidt, A., Yus, E., Aebersold, R., Herrmann, R., Böttcher, B., Frangakis, A. S., Russell, R. B., Serrano, L., Bork, P., and Gavin, A. C. (2009) Proteome organization in a genome-reduced bacterium. *Science*

- 326, 1235–1240
2. Zhang, Y., Thiele, I., Weekes, D., Li, Z., Jaroszewski, L., Ginalska, K., Deacon, A. M., Wooley, J., Lesley, S. A., Wilson, I. A., Palsson, B., Osterman, A., and Godzik, A. (2009) Three-dimensional structural view of the central metabolic network of *Thermotoga maritima*. *Science* **325**, 1544–1549
3. Kim, P. M., Lu, L. J., Xia, Y., and Gerstein, M. B. (2006) Relating three-dimensional structures to protein networks provides evolutionary insights. *Science* **314**, 1938–1941
4. Huang, Y. J., Hang, D., Lu, L. J., Tong, L., Gerstein, M. B., and Montelione, G. T. (2008) Targeting the human cancer pathway protein interaction network by structural genomics. *Mol. Cell. Proteomics* **7**, 2048–2060
5. Han, B. G., Dong, M., Liu, H., Camp, L., Geller, J., Singer, M., Hazen, T. C., Choi, M., Witkowska, H. E., Ball, D. A., Typke, D., Downing, K. H., Shatsky, M., Brenner, S. E., Chandonia, J. M., Biggin, M. D., and Glaeser, R. M. (2009) Survey of large protein complexes in *D. vulgaris* reveals great structural diversity. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 16580–16585
6. Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504
7. Cline, M. S., Smoot, M., Cerami, E., Kuchinsky, A., Landys, N., Workman, C., Christmas, R., Avila-Campillo, I., Creech, M., Gross, B., Hanspers, K., Isserlin, R., Kelley, R., Killcoyne, S., Lotia, S., Maere, S., Morris, J., Ono, K., Pavlovic, V., Pico, A. R., Vailaya, A., Wang, P. L., Adler, A., Conklin, B. R., Hood, L., Kuiper, M., Sander, C., Schmulevich, I., Schwikowski, B., Warner, G. J., Ideker, T., and Bader, G. D. (2007) Integration of biological networks and gene expression data using Cytoscape. *Nat. Protoc.* **2**, 2366–2382
8. Hu, Z., Mellor, J., Wu, J., Yamada, T., Holloway, D., and Delisi, C. (2005) VisANT: data-integrating visual framework for biological networks and modules. *Nucleic Acids Res.* **33**, W352–W357
9. Breitkreutz, B. J., Stark, C., and Tyers, M. (2002) Osprey: a network visualization system. *Genome Biol.* **3**, PREPRINT0012
10. Freeman, T. C., Goldovsky, L., Brosch, M., van Dongen, S., Mazière, P., Grocock, R. J., Freilich, S., Thornton, J., and Enright, A. J. (2007) Construction, visualisation, and clustering of transcription networks from microarray expression data. *PLoS Comput. Biol.* **3**, 2032–2042
11. Pavlopoulos, G. A., O'Donoghue, S. I., Satagopam, V. P., Soldatos, T. G., Pafilis, E., and Schneider, R. (2008) Arena3D: visualization of biological networks in 3D. *BMC Syst. Biol.* **2**, 104
12. Demir, E., Babur, O., Dogrusoz, U., Gursoy, A., Nisanci, G., Cetin-Atalay, R., and Ozturk, M. (2002) PATIKA: an integrated visual environment for collaborative construction and analysis of cellular pathways. *Bioinformatics* **18**, 996–1003
13. Nikitin, A., Egorov, S., Daraselia, N., and Mazo, I. (2003) Pathway studio—the analysis and navigation of molecular networks. *Bioinformatics* **19**, 2155–2157
14. Pavlopoulos, G. A., Wegener, A. L., and Schneider, R. (2008) A survey of visualization tools for biological network analysis. *BioData Min.* **1**, 12
15. DeLano, W. L. (2002) *The PyMOL Molecular Graphics System*, DeLano Scientific LLC, San Carlos, CA
16. Humphrey, W., Dalke, A., and Schulten, K. (1996) VMD: visual molecular dynamics. *J. Mol. Graph.* **14**, 33–38, 27–28
17. Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., and Ferrin, T. E. (2004) UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612
18. Murray-Rust, P., Rzepa, H. S., Williamson, M. J., and Willighagen, E. L. (2004) Chemical markup, XML, and the World Wide Web. 5. Applications of chemical metadata in RSS aggregators. *J. Chem. Inf. Comput. Sci.* **44**, 462–469
19. Sayle, R. A., and Milner-White, E. J. (1995) RASMOL: biomolecular graphics for all. *Trends Biochem. Sci.* **20**, 374
20. Moreland, J. L., Gramada, A., Buzko, O. V., Zhang, Q., and Bourne, P. E. (2005) The Molecular Biology Toolkit (MBT): a modular platform for developing molecular visualization applications. *BMC Bioinformatics* **6**, 21
21. Wang, Y., Geer, L. Y., Chappay, C., Kans, J. A., and Bryant, S. H. (2000) Cn3D: sequence and structure views for Entrez. *Trends Biochem. Sci.* **25**, 300–302
22. Jensen, L. J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., Doerks, T., Julien, P., Roth, A., Simonovic, M., Bork, P., and von Mering, C. (2009) STRING 8—a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.* **37**, D412–D416
23. Morris, J. H., Huang, C. C., Babbitt, P. C., and Ferrin, T. E. (2007) strucViz: linking Cytoscape and UCSF Chimera. *Bioinformatics* **23**, 2345–2347
24. Holland, E. C. (2000) Glioblastoma multiforme: the terminator. *Proc. Natl. Acad. Sci. U.S.A.* **97**, 6242–6244
25. Furnari, F. B., Fenton, T., Bachoo, R. M., Mukasa, A., Stommel, J. M., Stegh, A., Hahn, W. C., Ligon, K. L., Louis, D. N., Brennan, C., Chin, L., DePinho, R. A., and Cavenee, W. K. (2007) Malignant astrocytic glioma: genetics, biology, and paths to treatment. *Genes Dev.* **21**, 2683–2710
26. Atlas, T. C. G. (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061–1068
27. Parsons, D. W., Jones, S., Zhang, X., Lin, J. C., Leary, R. J., Angenendt, P., Mankoo, P., Carter, H., Siu, I. M., Gallia, G. L., Olivi, A., McLendon, R., Rasheed, B. A., Keir, S., Nikolskaya, T., Nikolsky, Y., Busam, D. A., Tekleab, H., Diaz, L. A., Jr., Hartigan, J., Smith, D. R., Strausberg, R. L., Marie, S. K., Shinjo, S. M., Yan, H., Riggins, G. J., Bigner, D. D., Karchin, R., Papadopoulos, N., Parmigiani, G., Vogelstein, B., Velculescu, V. E., and Kinzler, K. W. (2008) An integrated genomic analysis of human glioblastoma multiforme. *Science* **321**, 1807–1812
28. Kanehisa, M., and Goto, S. (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30
29. Aoki, K. F., and Kanehisa, M. (2005) Using the KEGG database resource. *Curr. Protoc. Bioinformatics* Chapter 1, Unit 1.12
30. Aoki-Kinoshita, K. F., and Kanehisa, M. (2007) Gene annotation and pathway mapping in KEGG. *Methods Mol. Biol.* **396**, 71–91
31. Matthews, L., Gopinath, G., Gillespie, M., Caudy, M., Croft, D., de Bono, B., Garapati, P., Hemish, J., Hermjakob, H., Jassal, B., Kanapin, A., Lewis, S., Mahajan, S., May, B., Schmidt, E., Vastrik, I., Wu, G., Birney, E., Stein, L., and D'Eustachio, P. (2009) Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res.* **37**, D619–D622
32. Karp, P. D., Ouzounis, C. A., Moore-Kochlacs, C., Goldovsky, L., Kaipa, P., Ahrén, D., Tsoka, S., Darzentas, N., Kunin, V., and López-Bigas, N. (2005) Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res.* **33**, 6083–6089
33. Schaefer, C. F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., and Buetow, K. H. (2009) PID: the Pathway Interaction Database. *Nucleic Acids Res.* **37**, D674–D679
34. Pico, A. R., Kelder, T., van Iersel, M. P., Hanspers, K., Conklin, B. R., and Evelo, C. (2008) WikiPathways: pathway editing for the people. *PLoS Biol.* **6**, e184
35. Kelder, T., Pico, A. R., Hanspers, K., van Iersel, M. P., Evelo, C., and Conklin, B. R. (2009) Mining biological pathways using WikiPathways web services. *PLoS One* **4**, e6447
36. Prasad, T. S., Kandasamy, K., and Pandey, A. (2009) Human Protein Reference Database and Human Proteinpedia as discovery tools for systems biology. *Methods Mol. Biol.* **577**, 67–79
37. Zhang, F., Gu, W., Hurler, M. E., and Lupski, J. R. (2009) Copy number variation in human health, disease, and evolution. *Annu. Rev. Genomics Hum. Genet.* **10**, 451–481
38. Horan, M. P. (2009) Application of serial analysis of gene expression to the study of human genetic disease. *Hum. Genet.* **126**, 605–614
39. Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U.S.A.* **95**, 14863–14868
40. Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Jr., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., and Tasumi, M. (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542
41. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000) The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242
42. Turner, N., and Grose, R. (2010) Fibroblast growth factor signalling: from development to cancer. *Nat. Rev. Cancer* **10**, 116–129
43. Rand, V., Huang, J., Stockwell, T., Ferriera, S., Buzko, O., Levy, S., Busam, D., Li, K., Edwards, J. B., Eberhart, C., Murphy, K. M., Tsiamouri, A., Beeson, K., Simpson, A. J., Venter, J. C., Riggins, G. J., and Strausberg, R. L. (2005) Sequence survey of receptor tyrosine kinases reveals mutations in glioblastomas. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 14344–14349

44. Lew, E. D., Furdai, C. M., Anderson, K. S., and Schlessinger, J. (2009) The precise sequence of FGF receptor autophosphorylation is kinetically driven and is disrupted by oncogenic mutations. *Sci. Signal.* **2**, ra6
45. Tsai, J., Lee, J. T., Wang, W., Zhang, J., Cho, H., Mamo, S., Bremer, R., Gillette, S., Kong, J., Haass, N. K., Sproesser, K., Li, L., Smalley, K. S., Fong, D., Zhu, Y. L., Marimuthu, A., Nguyen, H., Lam, B., Liu, J., Cheung, I., Rice, J., Suzuki, Y., Luu, C., Settachatgul, C., Shellooe, R., Cantwell, J., Kim, S. H., Schlessinger, J., Zhang, K. Y., West, B. L., Powell, B., Habets, G., Zhang, C., Ibrahim, P. N., Hirth, P., Artis, D. R., Herlyn, M., and Bollag, G. (2008) Discovery of a selective inhibitor of oncogenic B-Raf kinase with potent antitumor activity. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 3041–3046
46. Bae, J. H., Lew, E. D., Yuzawa, S., Tomé, F., Lax, I., and Schlessinger, J. (2009) The selectivity of receptor tyrosine kinase signaling is controlled by a secondary SH2 domain binding site. *Cell* **138**, 514–524
47. Meng, E. C., Pettersen, E. F., Couch, G. S., Huang, C. C., and Ferrin, T. E. (2006) Tools for integrated sequence-structure analysis with UCSF Chimera. *BMC Bioinformatics* **7**, 339
48. Xu, X., Zhao, J., Xu, Z., Peng, B., Huang, Q., Arnold, E., and Ding, J. (2004) Structures of human cytosolic NADP-dependent isocitrate dehydrogenase reveal a novel self-regulatory mechanism of activity. *J. Biol. Chem.* **279**, 33946–33957
49. Zhao, S., Lin, Y., Xu, W., Jiang, W., Zha, Z., Wang, P., Yu, W., Li, Z., Gong, L., Peng, Y., Ding, J., Lei, Q., Guan, K. L., and Xiong, Y. (2009) Glioma-derived mutations in IDH1 dominantly inhibit IDH1 catalytic activity and induce HIF-1 α . *Science* **324**, 261–265
50. Dang, L., White, D. W., Gross, S., Bennett, B. D., Bittinger, M. A., Driggers, E. M., Fantin, V. R., Jang, H. G., Jin, S., Keenan, M. C., Marks, K. M., Prins, R. M., Ward, P. S., Yen, K. E., Liao, L. M., Rabinowitz, J. D., Cantley, L. C., Thompson, C. B., Vander Heiden, M. G., and Su, S. M. (2009) Cancer-associated IDH1 mutations produce 2-hydroxyglutarate. *Nature* **462**, 739–744
51. Keiser, M. J., Setola, V., Irwin, J. J., Laggner, C., Abbas, A. I., Hufeisen, S. J., Jensen, N. H., Kuijter, M. B., Matos, R. C., Tran, T. B., Whaley, R., Glennon, R. A., Hert, J., Thomas, K. L., Edwards, D. D., Shoichet, B. K., and Roth, B. L. (2009) Predicting new molecular targets for known drugs. *Nature* **462**, 175–181
52. Weininger, D. (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28**, 31–36
53. Keiser, M. J., Roth, B. L., Armbruster, B. N., Ernsberger, P., Irwin, J. J., and Shoichet, B. K. (2007) Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.* **25**, 197–206
54. Pangalos, M. N., Neefs, J. M., Somers, M., Verhasselt, P., Bekkers, M., van der Helm, L., Fraiponts, E., Ashton, D., and Gordon, R. D. (1999) Isolation and expression of novel human glutamate carboxypeptidases with N-acetylated alpha-linked acidic dipeptidase and dipeptidyl peptidase IV activity. *J. Biol. Chem.* **274**, 8470–8483
55. Liang, R., Fei, Y. J., Prasad, P. D., Ramamoorthy, S., Han, H., Yang-Feng, T. L., Hediger, M. A., Ganapathy, V., and Leibach, F. H. (1995) Human intestinal H⁺/peptide cotransporter. Cloning, functional expression, and chromosomal localization. *J. Biol. Chem.* **270**, 6456–6463
56. Reaux, A., Iturrioz, X., Vazeux, G., Fournie-Zaluski, M. C., David, C., Roques, B. P., Corvol, P., and Llorens-Cortes, C. (2000) Aminopeptidase A, which generates one of the main effector peptides of the brain renin-angiotensin system, angiotensin III, has a key role in central control of arterial blood pressure. *Biochem. Soc. Trans.* **28**, 435–440
57. Juillerat-Jeanneret, L., Lohm, S., Hamou, M. F., and Pinet, F. (2000) Regulation of aminopeptidase A in human brain tumor vasculature: evidence for a role of transforming growth factor-beta. *Lab. Invest.* **80**, 973–980
58. Nanus, D. M., Engelstein, D., Gastl, G. A., Gluck, L., Vidal, M. J., Morrison, M., Finstad, C. L., Bander, N. H., and Albino, A. P. (1993) Molecular cloning of the human kidney differentiation antigen gp160: human aminopeptidase A. *Proc. Natl. Acad. Sci. U.S.A.* **90**, 7069–7073
59. Lasker, K., Topf, M., Sali, A., and Wolfson, H. J. (2009) Inferential optimization for simultaneous fitting of multiple components into a cryoEM map of their assembly. *J. Mol. Biol.* **388**, 180–194
60. Alber, F., Förster, F., Korkin, D., Topf, M., and Sali, A. (2008) Integrating diverse data for structure determination of macromolecular assemblies. *Annu. Rev. Biochem.* **77**, 443–477
61. Pintilie, G. D., Zhang, J., Goddard, T. D., and Gossard, D. C. (2010) Quantitative analysis of cryo-EM density map segmentation by watershed and scale-space filtering, and fitting of structures by alignment to regions. *J. Struct. Biol.* **170**, 427–438
62. Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., Moore, L., Adams, S. L., Millar, A., Taylor, P., Bennett, K., Boutillier, K., Yang, L., Wolting, C., Donaldson, I., Schandorff, S., Shewnarane, J., Vo, M., Taggart, J., Goudreaux, M., Muskut, B., Alfarano, C., Dewar, D., Lin, Z., Michalickova, K., Willems, A. R., Sassi, H., Nielsen, P. A., Rasmussen, K. J., Andersen, J. R., Johansen, L. E., Hansen, L. H., Jespersen, H., Podtelejnikov, A., Nielsen, E., Crawford, J., Poulsen, V., Sørensen, B. D., Matthiesen, J., Hendrickson, R. C., Gleeson, F., Pawson, T., Moran, M. F., Durocher, D., Mann, M., Hogue, C. W., Figeys, D., and Tyers, M. (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415**, 180–183
63. Gavin, A. C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L. J., Bastuck, S., Dümpelfeld, B., Edelmann, A., Heutier, M. A., Hoffman, V., Hoefert, C., Klein, K., Hudak, M., Michon, A. M., Schelder, M., Schirle, M., Remor, M., Rudi, T., Hooper, S., Bauer, A., Bouwmeester, T., Casari, G., Drewes, G., Neubauer, G., Rick, J. M., Kuster, B., Bork, P., Russell, R. B., and Superti-Furga, G. (2006) Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440**, 631–636
64. Krogan, N. J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A. P., Punna, T., Peregrín-Alvarez, J. M., Shales, M., Zhang, X., Davey, M., Robinson, M. D., Paccanaro, A., Bray, J. E., Sheung, A., Beattie, B., Richards, D. P., Canadien, V., Lalev, A., Mena, F., Wong, P., Starostine, A., Canete, M. M., Vlasblom, J., Wu, S., Orsi, C., Collins, S. R., Chandran, S., Haw, R., Ristone, J. J., Gandi, K., Thompson, N. J., Musso, G., St Onge, P., Ghanny, S., Lam, M. H., Butland, G., Altaf-Ul, A. M., Kanaya, S., Shilatifard, A., O'Shea, E., Weissman, J. S., Ingles, C. J., Hughes, T. R., Parkinson, J., Gerstein, M., Wodak, S. J., Emili, A., and Greenblatt, J. F. (2006) Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* **440**, 637–643
65. Collins, S. R., Kemmeren, P., Zhao, X. C., Greenblatt, J. F., Spencer, F., Holstege, F. C., Weissman, J. S., and Krogan, N. J. (2007) Toward a comprehensive atlas of the physical interactome of *Saccharomyces cerevisiae*. *Mol. Cell. Proteomics* **6**, 439–450
66. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29
67. van Dongen, S. (2000) *Graph Clustering by Flow Simulation*. Ph.D. thesis, University of Utrecht
68. Dengli, S., and Cramer, P. (2009) Torpedo nuclease Rat1 is insufficient to terminate RNA polymerase II in vitro. *J. Biol. Chem.* **284**, 21270–21279
69. Orlicky, S. M., Tran, P. T., Sayre, M. H., and Edwards, A. M. (2001) Dissociable Rpb4-Rpb7 subassembly of RNA polymerase II binds to single-strand nucleic acid and mediates a post-recruitment step in transcription initiation. *J. Biol. Chem.* **276**, 10097–10102
70. Edwards, A. M., Kane, C. M., Young, R. A., and Kornberg, R. D. (1991) Two dissociable subunits of yeast RNA polymerase II stimulate the initiation of transcription at a promoter in vitro. *J. Biol. Chem.* **266**, 71–75
71. Goddard, T. D., Huang, C. C., and Ferrin, T. E. (2007) Visualizing density maps with UCSF Chimera. *J. Struct. Biol.* **157**, 281–287
72. Kostek, S. A., Grob, P., De Carlo, S., Lipscomb, J. S., Garczarek, F., and Nogales, E. (2006) Molecular architecture and conformational flexibility of human RNA polymerase II. *Structure* **14**, 1691–1700
73. Tagari, M., Newman, R., Chagoyan, M., Carazo, J. M., and Henrick, K. (2002) New electron microscopy database and deposition system. *Trends Biochem. Sci.* **27**, 589
74. Meka, H., Werner, F., Cordell, S. C., Onesti, S., and Brick, P. (2005) Crystal structure and RNA binding of the Rpb4/Rpb7 subunits of human RNA polymerase II. *Nucleic Acids Res.* **33**, 6435–6444
75. Lasker, K., Phillips, J. L., Russel, D., Velazquez-Muriel, J., Schneidman-Duhovny, D., Tjioe, E., Webb, B., Schlessinger, A., and Sali, A. (2010) Integrative structure modeling of macromolecular assemblies from proteomics data. *Mol. Cell. Proteomics* **9**, 1689–1702
76. Förster, F., Lasker, K., Nickell, S., Sali, A., and Baumeister, W. (2010) Toward an integrated structural model of the 26 S proteasome. *Mol. Cell. Proteomics* **9**, 1666–1677

