Interactome3D: adding structural details to protein networks

Roberto Mosca[1], Arnaud Céol[1] & Patrick Aloy[1,2]

Network-centered approaches are increasingly used to understand the fundamentals of biology. However, the molecular details contained in the interaction networks, often necessary to understand cellular processes, are very limited, and the experimental difficulties surrounding the determination of protein complex structures make computational modeling techniques paramount. Here we present Interactome3D, a resource for the structural annotation and modeling of protein-protein interactions. Through the integration of interaction data from the main pathway repositories, we provide structural details at atomic resolution for over 12,000 protein-protein interactions in eight model organisms. Unlike static databases, Interactome3D also allows biologists to upload newly discovered interactions and pathways in any species, select the best combination of structural templates and build three-dimensional models in a fully automated manner. Finally, we illustrate the value of Interactome3D through the structural annotation of the complement cascade pathway, rationalizing a potential common mechanism of action suggested for several disease-causing mutations.

Proteins are effector molecules with key roles in virtually all events taking place within and between cells. However, they seldom act in isolation, and often biological processes are carried out by large molecular machines whose action is coordinated through complex networks of transient protein interactions. It is thus the inter-relationships between molecules, rather than the individual components, that ultimately determine the behavior of a biological system. Consequently, network-centered approaches are increasingly used to understand the fundamentals of biology and are becoming the starting point of many exciting new scientific developments[1,2].

Much effort has been devoted to systematically charting protein inter-relationships in several model organisms, including humans[3–5]. The valuable information contained in these high-throughput interactome data sets enables analyses of global properties of the systems. However, although interaction-discovery experiments can show that two proteins interact, they do not reveal the underlying molecular mechanisms that promote binding. A full understanding of how proteins physically interact can come only from high-resolution three-dimensional (3D) structures. Indeed, knowledge of the molecular details of binding has permitted a more rational analysis of exposed disease-causing mutations[6,7]; the design of experiments to disrupt an interaction and perturb systems where interactions are involved[8]; and distinction between 'singlish' and multi-interface hubs, providing hints for understanding network dynamics[9].

The exponential growth in the number of 3D structures stored in the Protein Data Bank (PDB)[10] means that it is increasingly rare to find a single protein for which no structural information is either available experimentally or readily accessible by straightforward homology modeling[11]. Nearly complete structural pictures for entire processes and organisms are likely to soon become available[12]. However, owing to the difficulty in obtaining high-resolution 3D information for protein-protein interactions, such data is very scarce. Thus, structural biology is somewhat lagging behind the new trends in high-throughput biology. In fact, there is a growing gap between the number of identified interactions and the number of interactions for which the 3D structure is known[13,14]. It is therefore crucial to devise effective strategies for incorporating structural information into interactome networks.

We have previously shown that homologous pairs of interacting proteins (that is, interologues[15]) tend to interact using the same binding interfaces[16], which has opened the door for homology modeling of protein-protein interactions[17]. This effectively means that, as for individual proteins, we can use the high-resolution 3D structure of a given protein-protein interaction as a template to model all the interactions that involve homologous proteins and for which the binding has been experimentally confirmed. The idea of increasing the structural coverage of the interaction space with computational models is highly attractive[18], and there have been considerable efforts toward the development of reliable methods to build such models[14,19]. In parallel, interaction data is also systematically collected and catalogued in various databases[20,21], whose contents are not yet unified[22]. Thus, although researchers have modeling algorithms as well as access to structural and interaction data, including databases that compile lists of potential domain-domain interaction templates[6], the information is scattered and difficult to

[1]Joint IRB-BSC Program in Computational Biology, Institute for Research in Biomedicine, Barcelona, Spain. [2]Institució Catalana de Recerca i Estudis Avançats, Barcelona, Spain. Correspondence should be addressed to P.A. (patrick.aloy@irbbarcelona.org).

RECEIVED 18 JULY; ACCEPTED 30 OCTOBER; PUBLISHED ONLINE 16 DECEMBER 2012; DOI:10.1038/NMETH.2289

integrate for the nonspecialist, which hampers its potential. Indeed, selection of the best template for modeling still requires manual intervention and technical expertise, which prevents the systematic application of modeling to newly discovered sets of interactions.

Here we present Interactome3D, a resource for the structural annotation and modeling of protein-protein interactions. Through the integration of interaction data contained in nine different databases and our homology modeling pipeline, we provide structural details for over 12,000 protein-protein interactions in eight model organisms, ranging from *Escherichia coli* to yeast to human. More importantly, we allow biologists to upload newly discovered interactions and pathways in any species, and to build 3D models in a fully automated manner. The resource has been implemented in a user-friendly and intuitive format, and can be found at http://interactome3d.irbbarcelona.org/.

RESULTS

Modeling strategy

Interactome3D is based on a fully automated computational approach, which is able to cope with large-scale data by taking advantage of parallel computing infrastructures (**Fig. 1**). The modeling pipeline can handle two types of input data: a set of interactions provided by the user, or a list of organisms for the modeling of either their entire interactomes or functional subparts of the interactomes. In contrast to the static nature of most available databases, which provide lists of structural templates for interactions[23–26] or map them to protein-protein interaction data sets[6], Interactome3D is a dynamic resource that collects all the necessary structural information for single proteins and binary interactions, automatically selects the best templates for modeling, and, where possible, applies a modeling pipeline and returns 3D coordinates of the binary complexes.

The first pipeline step is to collect structures for each individual protein in the set. We first identify the available experimental structures in the PDB and increase the structural coverage of the protein space by using high-quality homology models[11]. We then classify all the individual proteins into three categories: complete experimental structures (covering >80% of the length of the protein with 100% sequence identity), complete homology models (>80% coverage) and partial experimental structures or models (the rest). For the last category, the fragments are grouped together to cover the greatest possible length of the protein.

We then proceed to identify experimentally determined structures of each interaction or, when these are not available, suitable templates to model them. We consider all pairs of contacting proteins sharing over 30% sequence identity with the protein pairs to be modeled (target interaction) and apply a battery of filtering criteria (see Online Methods). Those potential templates fulfilling the requirements are then sorted with a scoring function that considers the completeness of the template and the sequence identity to the target. To improve structural coverage, we also include in the search partial templates involving structural domains in the two partners[23]. Interactome3D often identifies several domain-domain pairs on which a given interaction can be modeled. In those cases we select the best templates on the basis of domain-domain interaction preferences observed in the PDB[27]. The final models are built with Modeller[28], and the resulting structures at atomic resolution are checked for the presence of structural knots[29]. We finally rank all the structures and models on the basis of their completeness and quality, so that it is possible to

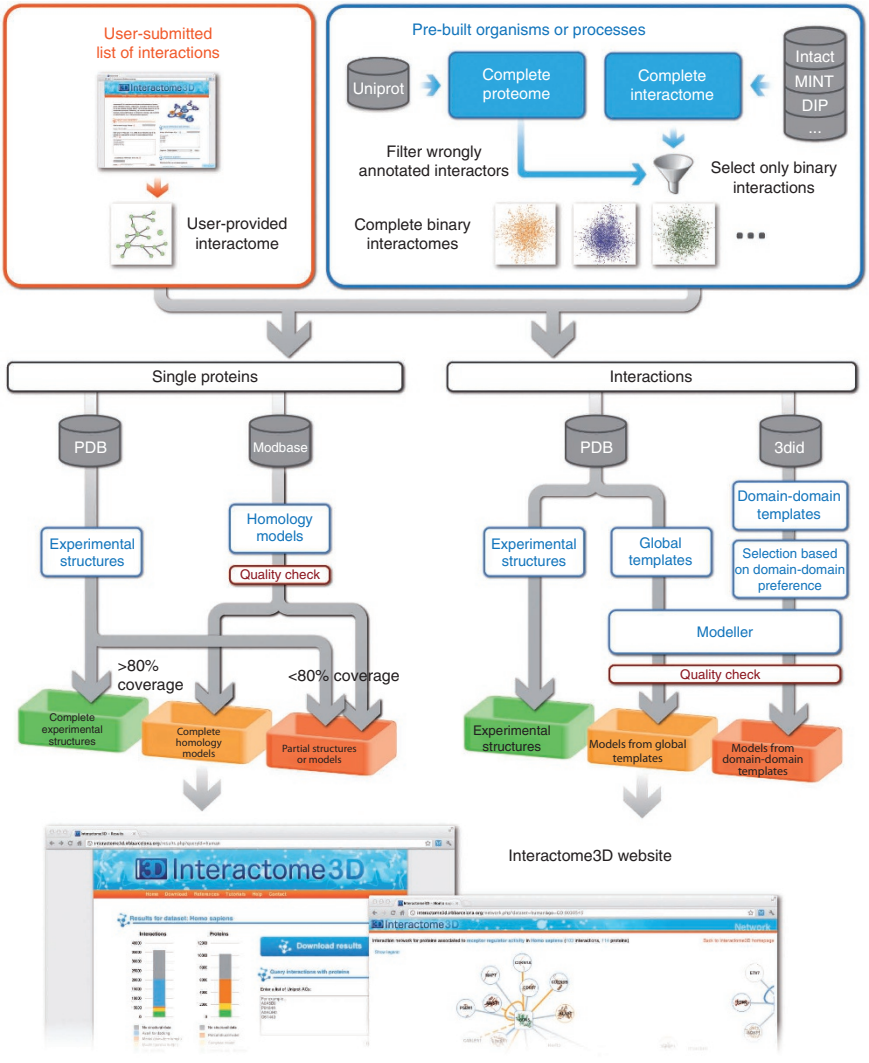

Figure 1 | The Interactome3D pipeline. The pipeline starts from either a user-provided list of interactions or a list of organisms for which it automatically compiles the most complete binary interactome. Experimental structures are collected and homology models built for both single interactors and binary complexes. Experimental structures are collected from the PDB, and homology models for single proteins are downloaded from Modbase. Models for interactions are built using Modeller starting from either global templates from the PDB or domain-domain templates from 3did.

Table 1 | Structural data collected by Interactome3D for eight organisms

	A. thaliana	*C. elegans*	*D. melanogaster*	*E. coli*	*H. pylori*	*H. sapiens*	*M. musculus*	*S. cerevisiae*
Proteins in proteome:	32,572	22,567	17,541	4,313	1,568	20,316	46,407	6,627
with complete structure	156	46	59	1,046	96	1,375	482	521
with complete model	2,331	2,209	2,343	1,422	572	2,484	4,870	926
with partial models	2,963	2,211	2,539	266	65	5,543	6,628	957
without structural data	27,122	18,101	12,600	1,579	835	10,914	34,427	4,223
Proteins in interactome:	4,362	4,130	6,736	2,234	779	10,147	2,966	5,589
with complete structure	107	37	50	935	86	1,224	261	503
with complete model	523	690	1,091	646	292	1,054	319	817
with partial structure	858	707	1,258	187	51	3,861	1,478	915
without structural data	2,874	2,696	4,337	466	350	4,008	908	3,354
Interactions:	10,368	7,450	20,294	7,406	1,614	35,948	4,225	24,033
with structure	140	45	116	963	88	3,181	622	822
with global model	795	247	314	79	28	1,730	512	385
with domain-domain model	448	106	250	31	9	709	168	379
with structures for interactors	974	1,188	2,748	4,481	359	14,861	1,400	5,651
without structural data	8,011	5,864	16,866	1,852	1,130	15,467	1,523	16,796

Represented are the proteomes and interactomes for the eight organisms and the number of proteins and interactions for which it was possible to collect experimental structures and build homology models.

obtain a representative set of models for each given interactome by selecting only the top-ranked models for every interaction.

Structured interactomes for eight model organisms

To assess the current coverage of Interactome3D, we applied it to eight model organisms: *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Escherichia coli*, *Drosophila melanogaster*, *Helicobacter pylori*, *Homo sapiens*, *Mus musculus* and *Saccharomyces cerevisiae* (**Table 1** and **Fig. 2**). For every organism, the pipeline first compiled the complete proteomes and then collected and merged binary interactions from several public databases (see Online Methods). Although the structural coverage for single proteins is extensive, the coverage of binary interactions is still limited and differs considerably among organisms. On average, there are structural data for 35% of each proteome; the human proteome has 46% structural coverage, and the greatest coverage is in *E. coli* (63%). The structural coverage of individual interacting proteins (proteins that are involved in at least one interaction) is slightly higher (**Supplementary Fig. 1**), ranging between 34% (for *A. thaliana*) and 79% (for *M. musculus*). As we expected, the percentage of binary interactions having structural data is much lower, ranging from a minimum of 3–5% for *C. elegans* and *D. melanogaster* to a maximum of 31% for *M. musculus*. Nevertheless, for some organisms there are thousands of structures and models available. For example, for humans there is structural data for more than 5,600 interactions (16% of the binary interactome), with models accounting for more than 2,400 interactions (7%). Given current estimates of the size of the human interactome (around 130,000 binary interactions[30]), a rough calculation suggests that, without the addition of any novel structural templates, Interactome3D should be able to provide structural details for ~16,000 interactions yet to be discovered (see Online Methods). Moreover, every template allows the modeling of four interactions (that is, ~6,000 human interactions are modeled using ~1,400 templates), and each time a novel experimentally determined structure of an interaction is deposited in the PDB, the applicability of Interactome3D will grow.

For both individual proteins and binary interactions, a large fraction of the available structures have been built by homology. These account for over 50% of the interactions having a structure, and up to 90% in *C. elegans* and *A. thaliana*. In addition, structural data are available for both partners for over 30,000 interactions in the eight species; in individual species the number ranges from almost 17,000 interactions in fly or yeast to 359 in *H. pylori*. These interactions are good candidates for high-throughput protein-protein docking[31], but owing to the much lower reliability of such docking methods with respect to comparative modeling methods, we have deliberately not incorporated docking data into Interactome3D. Overall, these figures show the potential of computational models to increase the structural coverage of the interaction space.

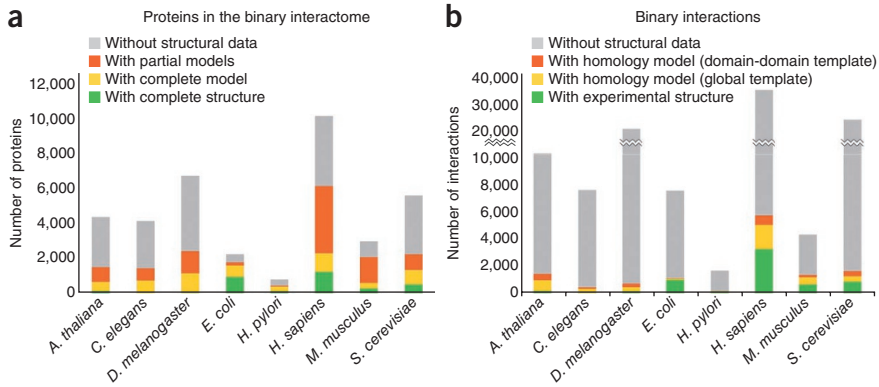

Figure 2 | Structural coverage of proteins and interactions for eight organisms. (**a**) Structural coverage of the single proteins in the binary interactomes. Coloring indicates the availability of complete experimental structures (green), complete models (yellow) or only partial models (orange). (**b**) Structural coverage of the binary interactions. Green, interactions for which an experimental structure is available; yellow and orange, interactions for which a homology model can be built from either a global template or a domain-domain template, respectively.

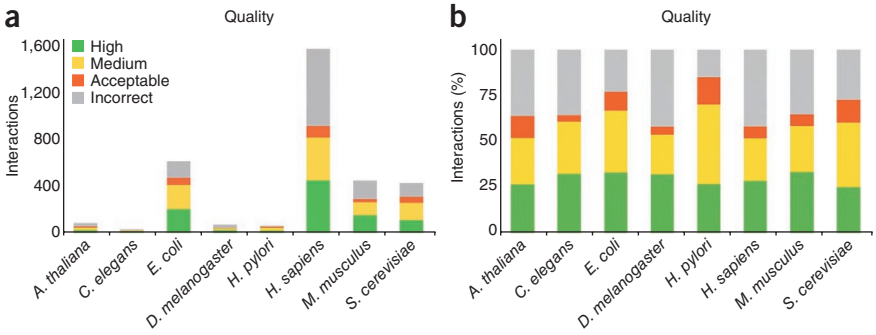

Figure 3 | Benchmarking of the homology models of interactions generated by Interactome3D. (**a**,**b**) Classification of homology models generated by Interactome3D for interactions already having an experimental structure, shown in absolute numbers (**a**) and as percentages within a given organism (**b**). High, medium, acceptable and incorrect correspond to the quality criteria used by CAPRI for the evaluation of protein-protein docking predictions.

Accuracy of the modeled interactions

To assess the quality of the structural models of binary protein-protein interactions produced by Interactome3D, we ran the pipeline for each of the eight organisms after removing all the experimental binary-interaction structures for the organism from the set of available templates. We then selected those interactions for which both a model can be produced and an experimental structure is available. We compared the models with the structures and classified the interaction models according to the criteria used to assess the results of the Critical Assessment of Prediction of Interactions (CAPRI; http://www.ebi.ac.uk/msd-srv/capri/)[32]. In brief, a model of an interaction is considered to be of high, medium or acceptable quality if, respectively, >50%, >30% or >10% of the native residue-residue contacts are conserved and the interface r.m.s. deviation (I-r.m.s. deviation) is <1.0 Å, 1.0 Å < I-r.m.s. deviation ≤ 2.0 Å or 2.0 Å < I-r.m.s. deviation ≤ 4.0 Å. Otherwise the models are considered incorrect. The number of interactions that could be evaluated ranged from 28 for *C. elegans* to 1,576 for *H. sapiens*, with a total of 3,297 interactions (**Supplementary Table 1**). Of these, 80% could be modeled on templates fetched from the PDB, and the remaining 20% were modeled on domain-domain structural templates from 3did (ref. 23).

For 64% of all interactions in Interactome3D there was at least one acceptable model for the interaction, and for 57% of interactions we produced a medium- to high-quality model (**Fig. 3**). We separated the models into those derived from global templates and those built from domain-domain templates, and found that, as expected, the average quality of the former was substantially higher (**Supplementary Fig. 2**). However, when we limited the evaluation to the top-ranked models, the results of the quality benchmarking did not change (**Supplementary Table 2** and **Supplementary Fig. 3**), demonstrating that the representative set of structures returned by Interactome3D indeed represents the most general set of results.

Toward structurally annotated pathways

Cellular networks and pathways provide a convenient visual summary of the results of hundreds of experiments devoted to charting the flow of signals or metabolites in a cell. Accordingly, much effort has been invested in the creation of systematically annotated pathway databases (for example, the Kyoto Encyclopedia of Genes and Genomes (KEGG)[33] or the Reactome[34]). However, although some of these resources try to capture specific details of the interaction (for example, phosphorylation sites), they generally lack functional details as to what an arrow between two proteins actually means. Interactome3D is, to our knowledge, unique in providing such structurally annotated pathways. The graphical interface of Interactome3D allows an intuitive visualization of a given pathway and provides information as to the pathway's structural coverage, offering hints for target-selection strategies in ongoing second-generation structural genomics initiatives[35].

We illustrate the value of Interactome3D here by annotating the complement cascade pathway from KEGG (**Fig. 4**). This pathway includes the interaction between complement factor H (CFH) and complement component 3 (C3) previously used in another study[6] to examine locus heterogeneity of disease-causing mutations. The example shows how Interactome3D provides context for the interaction and 3D coordinates, which permit detailed positioning of the mutated residues on the interaction interface.

The complement activation system is a component of the innate immune system that helps antibodies and phagocytic cells to protect the host against pathogens. Three pathways of complement activation are known: the classical pathway, the lectin pathway and the alternative pathway. Starting from the KEGG diagram (**Fig. 4**), we extracted a list of interactions and submitted them to Interactome3D. In total we retrieved complete experimental structures for 11 proteins, complete homology models for 4 and partial structures or models for 15 (**Supplementary Table 3**). There was only one protein (the C3a anaphylatoxin chemotactic receptor, C3AR1) for which the pipeline could not collect any structural data. Interactome3D was also able to collect experimental structures for six interactions corresponding to the complex between CFH and the complement C3d fragment (PDB 3OXU), the complex between the complement C3b fragment and complement factors B and D (PDB 2XWB), the C3 convertase C3bBb (PDB 2WIN), the subcomplex between the complement C5b fragment and complement component C6 (PDB 4E0S), and the interaction between the complement C3d fragment and complement receptor type 2 (PDB 3OED). For the other ten interactions the pipeline was able to produce homology models starting from global (three out of ten) or domain-domain structural templates (the remaining seven). They include the interaction between the C1 complex and complement factors 2 and 4, the interaction between mannan-binding lectin serine proteases 1 and 2 (MASP1 and MASP2) and complement C2 and C4, and the interaction between complement decay-accelerating factor (CD55) and C3.

With the structured version of the complement cascade pathway in hand, a straightforward application is the mapping and classification of disease-related mutations. In this case, mutations related to three different diseases have been described (complement components deficiency, complement factors deficiency and hemolytic uremic syndrome atypical). We were able to assign these mutations to specific proteins and interactions, and classify them as buried, surface and interface mutations (**Fig. 4**).

Indeed, the availability of structures and models of the interactions allowed us to identify not only whether a mutated residue is in the core or on the surface of a protein but also whether it lies on the binding interface between two proteins, potentially affecting interactions with specific partners. Examining their structural positions, we found that most of the mutations related to hemolytic uremic syndrome on CFH and C3, although they belong to two different proteins, are located on the same interaction interface (**Fig. 5**). The mapping of mutations within the sequence boundaries of interacting domains has previously suggested a common mechanism of action, namely alteration of the C3-CFH interaction[6]. However, the complete 3D coordinates of the interaction distinguish between mutations in the hydrophobic core of the interacting domains (for example, C1158W on C3) and

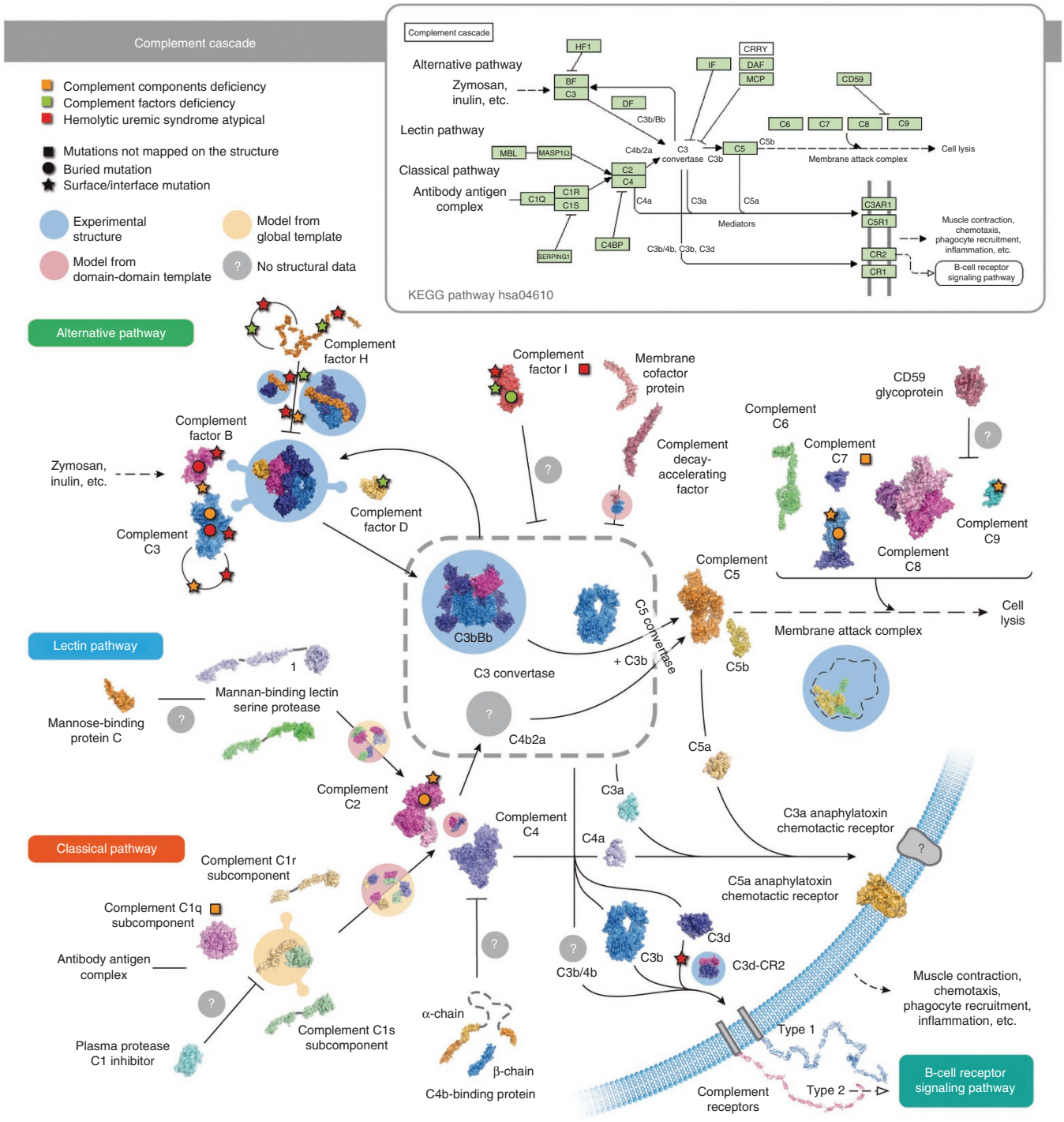

Figure 4 | Structural annotation of the complement cascade. Shown is the complement cascade pathway from KEGG (pathway ID hsa04610, inset) annotated with all the structural data collected by Interactome3D. For some of the proteins only partial models are available. We also mapped mutations related to three different diseases onto the collected structures and classified them as buried in the core of the proteins (circle), on the surface (star on the structure of a protein) or at an interface between two proteins (star on the line connecting two proteins). Mutations that could not be mapped onto the structure of the protein are represented with squares. Symbols of mutations are colored according to the disease to which they belong.

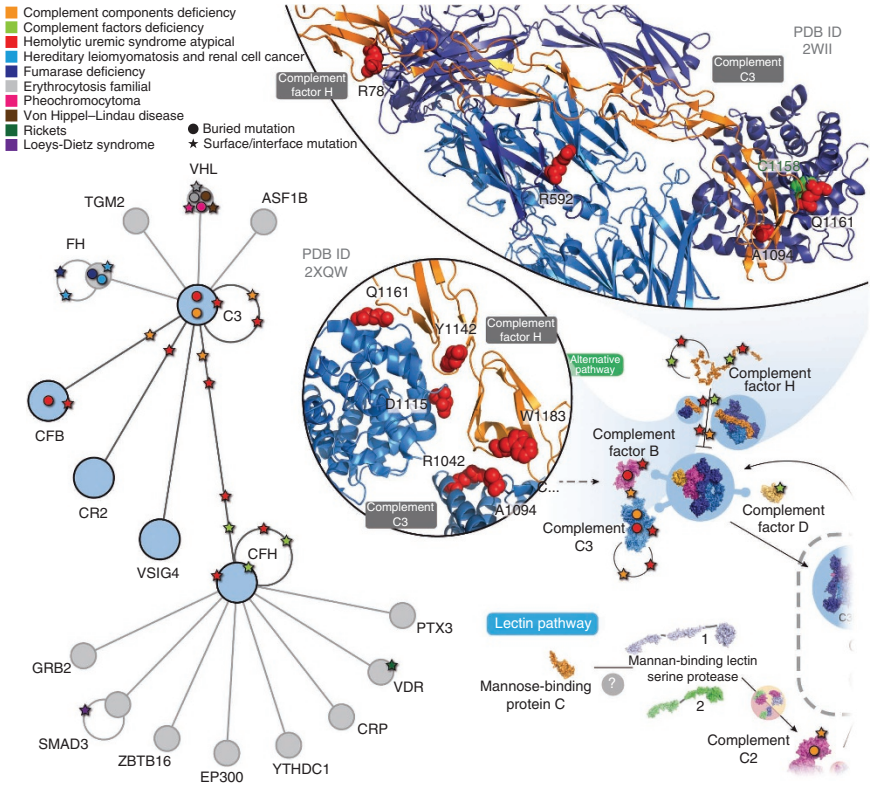

Figure 5 | Mapping disease mutations in the context of the structural interactome. Shown are complement C3 and CFH (right) and their interactions (left) either within the complement cascade (light blue) or taken from the human interactome (gray). Although VSIG4 is not in the KEGG pathway for the complement cascade, it is shown in blue because it is known to be a potent inhibitor of the complement pathway convertases[42,43]. Onto this interaction neighborhood we have mapped mutations for ten diseases (see key, top left) and classified them as either buried in the core of the protein (circles) or exposed on the surface (stars). Stars on a line between two proteins refer to mutations found at an interface. The two circular insets show the mapping of nine mutations related to hemolitic uremic syndrome atypical onto two different experimental structures (PDB 2XQW and 2WII) of the interaction between CFH (orange) and C3 (blue). Mutated residues are shown in atomic surface representation and colored red. Although the mutations belong to two different proteins, most are located on the same interaction interfaces, and these may all act by affecting the binding between the two proteins. One mutated residue (Cys1158, in green) belongs to the interacting domain A2M_comp of C3 and is buried inside the domain instead of being exposed on the interaction interface with CFH.

those on the contacting surfaces of the two proteins (for example, A1094V or Q1161K on C3 and R78G or Y1142D on CFH). Furthermore, the full coordinates place the mutation R1042L on C3 at the interface with CFH, even though the mutated residue is not inside the sequence boundaries of any described domain. Thus, Interactome3D provides complementary structural details for the hypothesis of a common mechanism for mutations causing hemolytic uremic syndrome, offering key structural details that could be used, for instance, to design small molecules able to stabilize the C3-CFH interface in the presence of the disease-causing mutations[36]. Furthermore, it is possible to contextualize mutations related to different diseases on the interaction neighborhood of CFH and C3 (**Fig. 5**), revealing the edgetic perturbations that the different diseases introduce and allowing the identification of potential relationships between them[37].

In addition to rationalizing disease-related mutations, structurally annotated pathways can also clarify the order of events in a signaling route by indicating which interactions are mutually exclusive owing to a common binding interface[18,38], suggest other factors affecting that order (for example, SH2 binding must be preceded by tyrosine phosphorylation) or provide a more rational basis for deciding where to interfere with a pathway in order to study it or to treat a particular disease[39].

The resource

The Interactome3D resource is available to the scientific community via a web interface (http://interactome3d.irbbarcelona.org/) that allows users to submit their own sets of interactions for processing. We make use of high-performance computing resources to annotate entire interactomes in very reasonable times (3,000 interactions per day). Interactome3D does not question the reliability of input interactions, and it is not meant to predict whether two proteins do or do not interact. Given a set of interactions, it collects all available structural data for them and builds reliable models for as many of the interactions as possible. Through the web interface, users can interactively visualize the network via CytoscapeWeb[40] and view detailed information for single proteins and binary complexes. Users can also query and browse the prebuilt data sets of entire organisms, using different functional criteria, and download them for offline analysis. Interactome3D will be updated regularly every 6 months to reflect the constantly increasing availability of protein interaction and structural data.

DISCUSSION

Protein interaction networks represent the cornerstone of much of modern biology. However, the level of detail provided by most 'omics' techniques is often insufficient to clearly link systemic and reductionist approaches based on cell or molecular biology. High-resolution 3D structural information could bridge this gap, but unfortunately, experimental bottlenecks surrounding structure determination make it currently unfeasible. Recent years have seen the emergence of pathway repositories and novel computational methods to complement experimental structures and increase the structural coverage of interactome networks and complexes[41], but they are often difficult for the nonspecialist to use. Interactome3D is a resource primarily for cell and molecular biologists, providing comprehensive, reliable and easy-to-interpret structural mapping of binary protein interactions that facilitates the integration of 3D data for protein interaction networks and cellular pathways. We hope that it will help biologists to get the most out of structural information and to routinely incorporate the added value

of atomic details into their experimental designs. In addition, the resource also permits downloading of all the data and generated models for offline large-scale analyses, which should enable a move away from artificial systemic representations and toward more realistic representations of cellular pathways. The challenge ahead is intimidating, but if successful, the structural annotation of cell pathways and interactome networks may provide crucial insights for the understanding of the complex genome-to-phenome relationships, both in health and disease.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. 3OXU, 2XWB, 2WIN, 4E0S, 3OED, 2WII, 2XQW.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

This work was partially supported by the Spanish Ministerio de Ciencia e Innovación (BIO2010-22073) and the European Commission under FP7 Grant Agreement 223101 (AntiPathoGN).

AUTHOR CONTRIBUTIONS

R.M. conceived and designed the work, wrote the manuscript, developed the pipeline, analyzed the data and implemented the Interactome3D web resource. A.C. compiled the integrated interaction database used by Interactome3D and implemented the Interactome3D web resource. P.A. conceived the work and wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/doifinder/10.1038/nmeth.2289.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Lee, M.J. *et al.* Sequential application of anticancer drugs enhances cell death by rewiring apoptotic signaling networks. *Cell* **149**, 780–794 (2012).
2. Shapira, S.D. *et al.* A physical and regulatory map of host-influenza interactions reveals pathways in H1N1 infection. *Cell* **139**, 1255–1267 (2009).
3. Rual, J.F. *et al.* Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173–1178 (2005).
4. Stelzl, U. *et al.* A human protein-protein interaction network: a resource for annotating the proteome. *Cell* **122**, 957–968 (2005).
5. Ewing, R.M. *et al.* Large-scale mapping of human protein-protein interactions by mass spectrometry. *Mol. Syst. Biol.* **3**, 89 (2007).
6. Wang, X. *et al.* Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nat. Biotechnol.* **30**, 159–164 (2012).
7. David, A., Razali, R., Wass, M.N. & Sternberg, M.J. Protein-protein interaction sites are hot spots for disease-associated nonsynonymous SNPs. *Hum. Mutat.* **33**, 359–363 (2012).
8. Dreze, M. *et al.* 'Edgetic' perturbation of a *C. elegans* BCL2 ortholog. *Nat. Methods* **6**, 843–849 (2009).
9. Kim, P.M., Lu, L.J., Xia, Y. & Gerstein, M.B. Relating three-dimensional structures to protein networks provides evolutionary insights. *Science* **314**, 1938–1941 (2006).
10. Berman, H.M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
11. Pieper, U. *et al.* ModBase, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Res.* **39**, D465–D474 (2011).
12. Zhang, Y. *et al.* Three-dimensional structural view of the central metabolic network of *Thermotoga maritima*. *Science* **325**, 1544–1549 (2009).
13. Pache, R.A. & Aloy, P. Incorporating high-throughput proteomics experiments into structural biology pipelines: identification of the low-hanging fruits. *Proteomics* **8**, 1959–1964 (2008).
14. Stein, A., Mosca, R. & Aloy, P. Three-dimensional modeling of protein interactions and complexes is going 'omics. *Curr. Opin. Struct. Biol.* **21**, 200–208 (2011).
15. Walhout, A.J. *et al.* Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science* **287**, 116–122 (2000).
16. Aloy, P., Ceulemans, H., Stark, A. & Russell, R.B. The relationship between sequence and interaction divergence in proteins. *J. Mol. Biol.* **332**, 989–998 (2003).
17. Aloy, P. *et al.* Structure-based assembly of protein complexes in yeast. *Science* **303**, 2026–2029 (2004).
18. Aloy, P. & Russell, R.B. Structural systems biology: modelling protein interactions. *Nat. Rev. Mol. Cell Biol.* **7**, 188–197 (2006).
19. Kuzu, G., Keskin, O., Gursoy, A. & Nussinov, R. Constructing structural networks of signaling pathways on the proteome scale. *Curr. Opin. Struct. Biol.* **22**, 367–377 (2012).
20. Kerrien, S. *et al.* The IntAct molecular interaction database in 2012. *Nucleic Acids Res.* **40**, D841–D846 (2012).
21. Licata, L. *et al.* MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res.* **40**, D857–D861 (2012).
22. Turinsky, A.L., Razick, S., Turner, B., Donaldson, I.M. & Wodak, S.J. Interaction databases on the same page. *Nat. Biotechnol.* **29**, 391–393 (2011).
23. Stein, A., Ceol, A. & Aloy, P. 3did: identification and classification of domain-based interactions of known three-dimensional structure. *Nucleic Acids Res.* **39**, D718–D723 (2011).
24. Davis, F.P. & Sali, A. PIBASE: a comprehensive database of structurally defined protein interfaces. *Bioinformatics* **21**, 1901–1907 (2005).
25. Gong, S. *et al.* PSIbase: a database of Protein Structural Interactome map (PSIMAP). *Bioinformatics* **21**, 2541–2543 (2005).
26. Finn, R.D., Marshall, M. & Bateman, A. iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics* **21**, 410–412 (2005).
27. Itzhaki, Z., Akiva, E. & Margalit, H. Preferential use of protein domain pairs as interaction mediators: order and transitivity. *Bioinformatics* **26**, 2564–2570 (2010).
28. Sali, A. & Blundell, T.L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779–815 (1993).
29. Taylor, W.R. A deeply knotted protein structure and how it might fold. *Nature* **406**, 916–919 (2000).
30. Venkatesan, K. *et al.* An empirical framework for binary interactome mapping. *Nat. Methods* **6**, 83–90 (2009).
31. Mosca, R., Pons, C., Fernandez-Recio, J. & Aloy, P. Pushing structural information into the yeast interactome by high-throughput protein docking experiments. *PLoS Comput. Biol.* **5**, e1000490 (2009).
32. Méndez, R., Leplae, R., De Maria, L. & Wodak, S.J. Assessment of blind predictions of protein-protein interactions: current status of docking methods. *Proteins* **52**, 51–67 (2003).
33. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* **40**, D109–D114 (2012).
34. Matthews, L. *et al.* Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res.* **37**, D619–D622 (2009).
35. Bravo, J. & Aloy, P. Target selection for complex structural genomics. *Curr. Opin. Struct. Biol.* **16**, 385–392 (2006).
36. Gordo, S. *et al.* Stability and structural recovery of the tetramerization domain of p53–R337H mutant induced by a designed templating ligand. *Proc. Natl. Acad. Sci. USA* **105**, 16426–16431 (2008).
37. Zhong, Q. *et al.* Edgetic perturbation models of human inherited disorders. *Mol. Syst. Biol.* **5**, 321 (2009).
38. Kiel, C. *et al.* Structural and functional protein network analyses predict novel signaling functions for rhodopsin. *Mol. Syst. Biol.* **7**, 551 (2011).
39. Russell, R.B. & Aloy, P. Targeting and tinkering with interaction networks. *Nat. Chem. Biol.* **4**, 666–673 (2008).
40. Lopes, C.T. *et al.* Cytoscape Web: an interactive web-based network browser. *Bioinformatics* **26**, 2347–2348 (2010).
41. Russel, D. *et al.* Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies. *PLoS Biol.* **10**, e1001244 (2012).
42. Vogt, L. *et al.* VSIG4, a B7 family-related protein, is a negative regulator of T cell activation. *J. Clin. Invest.* **116**, 2817–2826 (2006).
43. Wiesmann, C. *et al.* Structure of C3b in complex with CRIg gives insights into regulation of complement activation. *Nature* **444**, 217–220 (2006).

ONLINE METHODS

The Interactome3D pipeline. Interactome3D consists of an automated pipeline based on a series of sequential steps. The input to the pipeline can be either a set of interactions specified as a list of pairs of Uniprot accession codes, or a list of organisms specified by their NCBI Taxonomic IDs. In the latter case Interactome3D first compiles a complete proteome by retrieving all the Uniprot accession codes associated with the specified organism and annotated as 'Complete Proteome' (http://www.uniprot.org/taxonomy/complete-proteomes). For the human proteome, only the proteins annotated in SwissProt are considered. Interactome3D then collects a list of all the available experimentally identified binary protein-protein interactions for the organism from an integrated database. This database is generated by merging the data available from nine major public protein-interaction databases: Intact[20], MINT[21], DIP[44], MPIDB[45], MatrixDb[46], InnateDb[47], BioGRID[48], BIND (from the PSI-MI 2.5 version by Gary Bader's laboratory[49]) and HPRD[50]. To combine the interactions from different sources, Interactome3D remaps all protein references to Uniprot accession codes, using either the mapping provided by Uniprot and downloadable from the Uniprot FTP repository or the PICR web service provided by the European Bioinformatics Institute[51]. To avoid duplicated references, proteins with 100% sequence identity are merged using the Uniref100 database[52]. The pipeline rejects all entries for which no Uniprot reference is found, as well as those for which the mapping is ambiguous (the same protein can be mapped to more than one Uniprot accession code).

For large-scale experiments (more than 100 interactions detected) curated by more than one database, Interactome3D imports the corresponding interactions from only one of the sources, with the following prioritization: first from the IMEx databases (Intact, MINT, DIP, MatrixDb, Innatedb and MPIDB), then from BIND, BioGRID (which annotates proteins with gene identifiers) and finally HPRD (which provides the lowest level of annotation). The same prioritization is used for publications involving interactions between more than two proteins (complexes). For small-scale experiments curated by more than one database, the corresponding interactions in all the databases are merged.

The pipeline finally identifies the interactions for which there is reliable evidence that they are binary—that is, physical interactions between two individual proteins. These interactions are selected either with the detection method (see next section) or on the basis of the annotations provided by the curators (physical associations and direct interactions). This latter criterion is applied only to entries that are curated according to the IMEx[53] or MIMIx[54] standards—that is, these entries have an IMEx identifier or a flag is present that specifies that the curation depth is either IMEx, MIMIx or light curation[55]. These IMEx and MIMIx entries may include, in some cases, interactions identified in larger complexes for which additional experimental or computational validation has been done (and reported in the original publication) to infer the binary nature of the interaction[56]. Interactome3D excludes interactions for which the Uniprot accession code of at least one of the interactors does not belong to the previously compiled proteome, to avoid data that are potentially incomplete or outdated.

After collecting the binary interactome, Interactome3D searches for experimental structures and homology models of the single proteins in the proteome or in the list of interactions provided by the user. First, BLAST is used to search the database of sequences corresponding to protein chains in the PDB. We consider a protein chain in a PDB file to be an experimental structure for a protein if the alignment between the protein and the chain has 100% sequence identity and covers at least 95% of the length of the chain. To complete the search, Interactome3D also uses the mapping between Uniprot and PDB provided by the SIFTS initiative (http://pdbe.org/sifts)[57]. At the same time the pipeline downloads from Modbase[11] all the available homology models for the proteome and filters out all the models having a ModPipe Quality Score (MPQS) <1.1 or, if the MPQS is not available, models having a sequence identity lower than 30% and a ga341 score <0.7 (ref. 58). As some models in Modbase correspond to sequences from previous versions of Uniprot, Interactome3D confirms the match between the query sequence and the model.

Experimental structures and models are then merged and ranked in the following order: (i) complete experimental structures, ordered by decreasing coverage (a structure is considered to be complete if it covers more than 80% of the protein length); (ii) complete homology models, in decreasing order of sequence identity of the template to the target; (iii) partial experimental structures and homology models. For partial structures and models, Interactome3D creates groups of nonredundant structures that, together, cover the largest possible portion of the protein sequence. The groups are generated with a greedy algorithm that sorts the structures by decreasing values of the following scoring function:

$$Score1 = \alpha \cdot \mathrm{SeqId}/100 \cdot \mathrm{Cov}/100 + (1-\alpha) \cdot \mathrm{Cov}/100 \quad (1)$$

The scoring function takes into account the coverage of the structure (Cov, in the range of 0–100) and the sequence identity of the template to the target protein (SeqId, in the range of 0–100). The parameter α is used to balance the contribution of the two terms (sequence identity and coverage) and is selected on the basis of empirical observations (see equation (2) later in this section for further details). Interactome3D assigns the same rank to every group of structures and models, and orders the components according to their starting position in the sequence of the protein.

The next step is the collection of experimental structures for interactions. Interactome3D checks, for each interaction, whether there are experimental structures of the two interactors that correspond to chains belonging to the same PDB file. In this case it collects all available biological units from the PDB file and verifies whether, in any of the biological units, the two chains are in contact. The biological units used by Interactome3D are the ones provided by the PDB and available for download from the PDB FTP site (under the path /pub/pdb/data/biounit). We consider two chains as interacting if they have at least five residue-residue contacts, including (i) covalent interactions (disulfide bridges), defined as two sulfur atoms of a pair of cysteines at a distance ≤2.56 Å (two times the covalent radius of sulfur plus 0.5 Å); (ii) hydrogen bonds, defined as all atom pairs N-O and O-N at a distance ≤3.5 Å; (iii) salt bridges, defined as all atom pairs N-O and O-N at a distance ≤5.5 Å; and (iv) van der Waals interactions, defined as all pairs of carbon atoms at a distance ≤5.0 Å. Any pair of atoms at a distance less than the sum of the two covalent radii plus 0.5 Å that are not forming a disulfide bridge are considered clashes, and are not counted.

doi:10.1038/nmeth.2289

For all the interactions that do not have an experimental structure, Interactome3D searches for structural templates for homology modeling and filters them using the following criteria: (i) X-ray structures are preferred over NMR structures, (ii) structures with resolutions below 5 Å are prioritized and (iii) if the minimum sequence identity of a template with the two target proteins is ≤40%, the template is used only if it has an InterPreTS *z*-score ≥1.65 (that is, 95% confidence level)[59].

We rank the remaining templates (if any) according to the following scoring function:

$$\begin{aligned} Score2 = {} & \alpha \cdot \mathrm{SeqId}_1/100 \cdot \mathrm{Cov}_1/100 + (1-\alpha) \cdot \mathrm{Cov}_1/100 \\ & + \alpha \cdot \mathrm{SeqId}_2/100 \cdot \mathrm{Cov}_2/100 + (1-\alpha) \cdot \mathrm{Cov}_2/100 \end{aligned} \quad (2)$$

where SeqId_i is the sequence identity of the *i*th interactor to the corresponding PDB chain and Cov_i is its coverage. The parameter α is used to balance the contribution of the two terms (sequence identity and coverage) so that sequence identity is given priority over coverage, except when there are several models with similar sequence identity, in which case a higher coverage is preferred to a lower one. α is set to 0.95 on the basis of manual empirical testing with sets of interactions having similar sequence identity (within approximately 5% difference) but very different structural coverage (greater than 20% difference).

Interactome3D selects the top three ranked templates for modeling, and Modeller[28] is used for the modeling. In preparation for the modeling phase the target proteins are aligned to the template using the align2d command of Modeller. Interactome3D crops unaligned stretches of residues outside the boundaries of the template at the N and C termini of the target proteins (to ensure that only the template section of the protein is modeled). The pipeline produces five models for every template and checks them with KNOT[29] to identify structural knots. Modeller can sometimes produce knotted structures, particularly when there are long insertions in the target protein compared with the template (http://salilab.org/modeller/FAQ.html#14). Interactome3D then selects the model having the lowest DOPE score[60], as reported by Modeller, and not presenting knots in the main chain. To prevent disconnected models owing to deletions in the target proteins in the region corresponding to the binding interface in the template, Interactome3D further checks the final models for contacts between the two partners and discards them if they show less than five residue-residue contacts.

For those interactions that have neither experimental structures nor modeling templates according to the protocol described above, Interactome3D searches for domain-domain structural templates. First, Pfam domains[61] are assigned to both the interactors. Then, for every pair of potentially interacting domains, Interactome3D searches for the availability of a structural template in 3did[23]. For those domain-domain pairs having a template, Interactome3D applies a prioritization algorithm based on the preference relation described in a previous work[27]. For every selected domain-domain pair, all the corresponding templates are sorted by their score in equation (2), and the top-ranked template is selected for modeling. Modeling is performed with Modeller[28] using the protocol described above. In this case, the target proteins are aligned to the template using the Pfam HMM profiles of the matching domains and the hmmalign command from the HMMER suite[62]. For those interactions for which no experimental structure is available and Interactome3D cannot find either a global or a domain-domain structural template, the pipeline cannot perform any modeling and the interaction is categorized as not having any available structural data.

It should be noted that Interactome3D also contains, and it is able to model, domain-motif interactions, in which a globular domain in one protein recognizes a short linear motif in another[63]. However, we do not give special treatment to these interactions, and they are subject to the same strict criteria of sequence identity and coverage described above.

The projection on the number of interactions that could be structurally annotated by Interactome3D on a complete human binary interactome is done by considering that the current version of the human binary interactome contains around 36K interactions, of which we can currently model ~6,000 (~16%). Given that the estimated size of the human binary interactome is 130,000 (ref. 30), we can predict that 16% of the remaining ~100,000 interactions could be assigned an experimental structure or a homology model based on the current available structural data in the PDB, which yields the figure of ~16,000 interactions stated herein.

Experimental methods for the detection of binary interactions. The selection of binary interactions is based on the method by which the interactions have been detected. The list of experimental methods to detect binary interactions was updated from ref. 3 to include all the children terms in the current PSI MI 2.5 ontology[64]. The updated list includes the following methods: two-hybrid array, two-hybrid fragment pooling approach, two-hybrid pooling approach, two-hybrid lexa b52 complementation, lex-a dimerization assay, lexa dimerization assay, gal4 vp16 complementation, cytoplasmic complementation assay, β-galactosidase complementation, β-lactamase complementation, adenylate cyclase complementation, reverse ras recruitment system, dehydrofolate reductase reconstruction, ubiquitin reconstruction, green fluorescence protein complementation assay, protein kinase A complementation, cross-linking study, protein cross-linking with a bifunctional reagent, nucleic acid UV cross-linking assay, NMR, peptide array, protein array, protein *in situ* array, ping, proteinchip(r) on a surface-enhanced laser desorption/ionization, scintillation proximity assay, kinase scintillation proximity assay, surface plasmon resonance, surface plasmon resonance array, X-ray crystallography.

Structural mapping of the complement cascade pathway from KEGG. From the pathway for the complement and coagulation cascades in KEGG (pathway ID hsa04610) we extracted the part of the pathway corresponding to the complement cascade alone (**Fig. 4**). We generated a list of the protein-protein interactions in the pathway and submitted the list to Interactome3D. From the results returned by Interactome3D we selected, manually, a list of structures to be used for the structural annotation of the pathway (**Fig. 4** and **Supplementary Table 3**). We obtained disease-related mutations from the Uniprot humsavar list (Uniprot release 2011_11) and clustered the diseases into classes (**Supplementary Table 4**). To classify mutations as buried or surface mutations, we analyzed the structures for single proteins using NACCESS. We classified a residue as 'surface' if the accessible solvent area was at least 5% of the residue's total surface, and otherwise classified

it as 'buried'[65,66]. We further classified residues as belonging to an interface if upon binding they lose 1 Å^2 of accessible solvent area.

44. Salwinski, L. *et al.* The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.* **32**, D449–D451 (2004).
45. Goll, J. *et al.* MPIDB: the microbial protein interaction database. *Bioinformatics* **24**, 1743–1744 (2008).
46. Chautard, E., Fatoux-Ardore, M., Ballut, L., Thierry-Mieg, N. & Ricard-Blum, S. MatrixDB, the extracellular matrix interaction database. *Nucleic Acids Res.* **39**, D235–D240 (2011).
47. Lynn, D.J. *et al.* InnateDB: facilitating systems-level analyses of the mammalian innate immune response. *Mol. Syst. Biol.* **4**, 218 (2008).
48. Stark, C. *et al.* The BioGRID Interaction Database: 2011 update. *Nucleic Acids Res.* **39**, D698–D704 (2011).
49. Isserlin, R., El-Badrawi, R.A. & Bader, G.D. The Biomolecular Interaction Network Database in PSI-MI 2.5. *Database (Oxford)* **2011**, baq037 (2011).
50. Keshava Prasad, T.S. *et al.* Human Protein Reference Database–2009 update. *Nucleic Acids Res.* **37**, D767–D772 (2009).
51. Côté, R.G. *et al.* The Protein Identifier Cross-Referencing (PICR) service: reconciling protein identifiers across multiple source databases. *BMC Bioinformatics* **8**, 401 (2007).
52. UniProt Consortium. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* **40**, D71–D75 (2012).
53. Orchard, S. *et al.* Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nat. Methods* **9**, 345–350 (2012).
54. Orchard, S. *et al.* The minimum information required for reporting a molecular interaction experiment (MIMIx). *Nat. Biotechnol.* **25**, 894–898 (2007).
55. Ceol, A. *et al.* MINT, the molecular interaction database: 2009 update. *Nucleic Acids Res.* **38**, D532–D539 (2010).
56. Hu, P. *et al.* Global functional atlas of *Escherichia coli* encompassing previously uncharacterized proteins. *PLoS Biol.* **7**, e96 (2009).
57. Velankar, S. *et al.* E-MSD: an integrated data resource for bioinformatics. *Nucleic Acids Res.* **33**, D262–D265 (2005).
58. Eswar, N. *et al.* Tools for comparative protein structure modeling and analysis. *Nucleic Acids Res.* **31**, 3375–3380 (2003).
59. Aloy, P. & Russell, R.B. InterPreTS: protein interaction prediction through tertiary structure. *Bioinformatics* **19**, 161–162 (2003).
60. Shen, M.Y. & Sali, A. Statistical potential for assessment and prediction of protein structures. *Protein Sci.* **15**, 2507–2524 (2006).
61. Punta, M. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **40**, D290–D301 (2012).
62. Eddy, S.R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
63. Stein, A. & Aloy, P. Novel peptide-mediated interactions derived from high-resolution 3-dimensional structures. *PLoS Comput. Biol.* **6**, e1000789 (2010).
64. Kerrien, S. *et al.* Broadening the horizon–level 2.5 of the HUPO-PSI format for molecular interactions. *BMC Biol.* **5**, 44 (2007).
65. Jones, S., Marin, A. & Thornton, J.M. Protein domain interfaces: characterization and comparison with oligomeric protein interfaces. *Protein Eng.* **13**, 77–82 (2000).
66. Miller, S., Janin, J., Lesk, A.M. & Chothia, C. Interior and surface of monomeric proteins. *J. Mol. Biol.* **196**, 641–656 (1987).

doi:10.1038/nmeth.2289