Modeling and Analyzing Three-Dimensional Structures of Human Disease Proteins

Yuzhen Ye, Zhanwen Li, and Adam Godzik

Pacific Symposium on Biocomputing 11:439-450(2006)

MODELING AND ANALYZING THREE-DIMENSIONAL STRUCTURES OF HUMAN DISEASE PROTEINS

YUZHEN YE, ZHANWEN LI and ADAM GODZIK

Bioinformatics and Systems Biology Program, The Burnham Institute, La Jolla, CA 92037, USA

Three-dimensional structures of proteins, experimental or predicted, show us how these molecular machines actually work. With the help of information on disease-related mutations, they can also show us how they malfunction in diseases. Such understanding, currently lacking for most human diseases, is an important first step before designing drugs or therapies to cure specific diseases. Here we used homology modeling to model human disease-related proteins, and studied structural characteristics of disease related mutations and compared them with non synonymous SNPs. 1484 domains from 874 proteins were modeled, and together with experimentally determined structures of 369 domains they provided the structural coverage of 48% of total residues in 1237 human disease proteins. We found that disease-related mutations have statistically significantly preference to form clusters on protein surfaces. In contrast, the non-synonymous SNPs appear to be randomly distributed on the surface. We interpret these results as an indication that disease mutations affect protein-protein interaction interfaces. This interpretation is supported by the analysis of 8 experimentally determined complexes between disease proteins, where disease-related mutations are clearly located in the binding interface of proteins, while SNPs are not. The non-uniform distribution of disease mutations indicates that we can use this feature as guidance in modeling and evaluating human disease proteins and their complexes. We set up a resource for Disease Protein Models (DPM at http://ffas.burnham.org/DPM), which can be used for studying the relation between disease and mutation / polymorphism sites in the context of protein 3D structures and complexes.

1. Introduction

Disease-related proteins are of great research interest for both experimental and computational scientists. Their high value in medicine and human health stems from the fact that they provide molecular picture of disease processes, a necessary prerequisite to rational drug development. As of today, thousands of genes (proteins) have been identified to be associated with various diseases in humans. Most often, mutations in these proteins have been identified in patients suffering from a particular disease. Mutation data is often the first source allowing us to study these diseases on the molecular level. Independently, technological advances in large-scale genome sequencing has allowed us to study human genomic variation and identify large numbers of SNPs (Single Nucleotide Polymorphisms), some of which cause changes of amino acids in the protein product of a gene (i.e., non-synonymous SNPs, nsSNPs) [1]. SNPs

databases become an easy but valuable resource for studying genetic variations in human population. A vast majority of the nsSNPs has not been studied experimentally, and it is generally assumed that since nsSNPs are present in large sections of the population they are not strongly associated with diseases.

Many computational methods have been applied to study the effects of mutations (such as in P53 proteins [2]) and to predict the effects of the nsSNPs based on protein sequences, amino acid conservation and protein structures [3-6]. Structural information has been extensively used for studying the effects of mutations and nsSNPs [7] and a number of resources have been developed for mapping the SNPs onto the structures, such as MutDB (http://mutdb.org/) [8], SNPs3D (http://www.snps3d.org/), PolyPhen (http://www.bork.embl-heidelberg.de/PolyPhen/) [5] and SAAP (http://acrmwww.biochem.ucl.ac.uk/saap/) [2]. These resources are largely limited to human proteins with experimentally determined structures and there are still only a very small number of such proteins. Despite tremendous advances in recent years, experimental determination of protein structure is still time-consuming and expensive, especially for Eukaryotic proteins. It is possible to circumvent this limitation by use of comparative modeling, and this approach has been applied to disease proteins and used for SNPs annotation, including LS-SNP [9] and ModSNP [10].

In this work, we used distant homology recognition in conjunction with comparative modeling to build models of three-dimensional structures of human disease proteins. This strategy, validated in fold recognition test, can greatly increase the structural coverage of these proteins. Using the predicted structures, we further analyzed and compared the distribution of disease mutations and nsSNPs in the 3D space. The observation that spatial distribution of disease-related mutations is significantly different from that of usually benign nsSNPs suggests a possible explanation of different effects of the two groups. We hypothesize that disease mutations affect protein-protein interaction interfaces and thus disrupt functional networks within the cell. Detailed analyses of several available structures of experimentally determined complexes between disease proteins support this hypothesis.

2. Methods

2.1. *Data collection*

We focused on 1,237 human disease proteins and the corresponding mutation and SNPs information (discarding the variant sites marked as “unclassified”) as identified by SwissProt database (http://us.expasy.org/sprot/). Structures of

experimentally determined human disease proteins and their complexes were identified and downloaded from the Protein Database web site (http://www.rcsb.org).

2.2. *Homology modeling and model quality assessment*

We used FFAS [11], a profile–profile alignment and fold-recognition tool, for identifying the templates for modeling and generating the alignments. The alignments from FFAS were used as inputs for modeling packages Jackal [12] and Modeller [13], which were used to build three-dimensional models using the default options (Modeller models were used when no models can be produced by Jackal). All the models can be found at the DPM Web site at http://ffas.burnham.org/DPM and can be downloaded or viewed with a MDL Chime (http://www.mdl.com/products/framework/chime/) enabled browser.

Model quality was evaluated by the PSQS (Protein Structure Quality Score, http://www1.jcsg.org/psqs/psqs.cgi), an energy-like measure of quality of protein structures, calculated based on the statistical potentials of mean force describing interactions between residue pairs and between single residues and solvent, shown before to correlate well with model quality and accuracy [14]. Similar protocol is used for building molecular replacement templates in the Joint Center for Structural Genomics [15] and in several other large scale modeling projects (manuscript in preparation).

2.3. *Spatial distribution of disease-related mutants on the protein structures*

We define a residue as being in the core if its solvent accessible area is less than 5% of its maximum possible surface area in a fully extended conformation. By this definition about 25% of residues in an average protein are in the core. We used Lee & Richard method [16] to compute the atomic solvent accessible area of proteins.

We use the size of the *largest connected component* of the mutations graph as a measure of clustering of mutations, and its significance is calculated by a permutation test. For example, suppose a protein structure has N residues and M mutation sites. We first generate a graph of M nodes and link an edge between two nodes if they are within contact distance (i.e., minimum distance between any two atoms of the residues is $\leq$5.0Å). The graph is then partitioned into connected components [17]; and the size of the largest component is used as the clustering index of mutation sites. To compute the significance, we randomly select M residues out of N and do the same computation for R times. The significance of the clustering of the mutation positions is then defined as (the

number of permutations with clustering index ≥ the clustering index of mutation sites) / *R*. See Figure 1 for a schematic illustration.

2.4. *Analysis of protein complex structures*

In addition to the models of disease proteins, we analyzed experimental structures of complexes of disease proteins, if available. Residues were considered as being on the complex interface if their burial/exposed status between the complex and individual structures changed, here defined as the difference of their solvent accessible area in complex and in individual is less than a cutoff (e.g, 5Å^2).

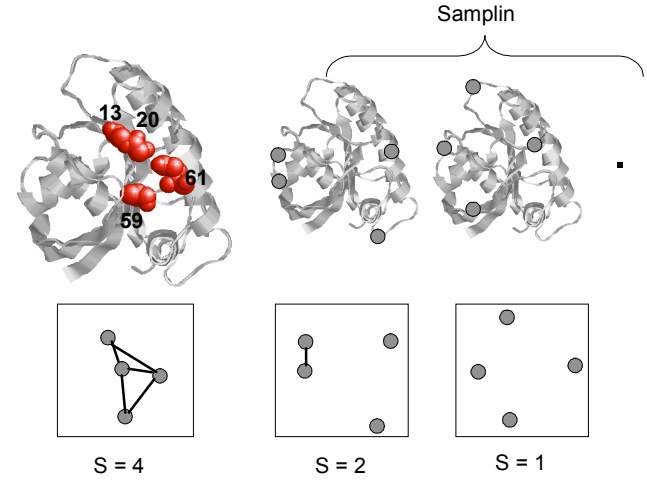

Figure 1. A schematic illustration of computing the significance of clustering of residues by a permutation test. Here 4 residues of position 13, 20, 59 and 61 highlighted in CPK are clustered, and the size of largest component of the graph with 4 nodes, S, is 4. The significance of the clustering P(S = 4) is (the number of samplings with S ≥4) / (total number of samplings).

3. Results

3.1. *Statistics of models and quality evaluation*

A total of 1,484 models were built for 874 proteins (many proteins are multi-domain and models were built for individual domains (if possible)). Together

with 369 experimentally determined structures (only one structure from all that cover the same or very similar region in a target protein was counted), 1859 structures in total cover at least partially 1,064 out of the 1,237 human disease proteins. The structural coverage counted as the percentage of residues that could be mapped to the structures is 9% by experimentally determined structures, 39% by models, and 48% in total. For example, alpha-2-macroglobulin receptor-associated protein (SwissProt accession number P30533) is 357 aa long, and only a 82 aa N-terminal fragment was determined experimentally (PDB code 1op1); a model covering 199 aa at its C-terminal domain can be built using 1bf5 chain A as a template. While much higher structural coverage than 48% was reported for bacterial genomes [18], eukaryotic genomes, such as human were expected to have lower coverage.

The quality of the homology model is determined by a combination of the performance of the modeling algorithm and the quality of the alignments. Therefore, the quality of our models is largely determined by the performance of FFAS used to detect templates and produce alignment for homology modeling. FFAS benchmarks have shown that predictions with scores lower than –9.5 (the cutoff used in this work) should have less than 3% of false positives [11], and that its alignment quality is significantly higher than PSI-BLAST alignments. An independent measure of a model quality can be provided by empirical energy parameters, such as for instance calculated by a PSQS server (http://www1.jcsg.org/psqs/psqs.cgi) [14]. 81% of the models have good overall PSQSs (Protein Structure Quality Score < 0) (see the DPM website for the detailed results). In addition we emphasize that features such as relative position of a residue on the surface or in the core of the protein tend to be well conserved even in relatively inaccurate models.

3.2. *Distribution of mutations (disease-related mutations and SNPs)*

A total of 6,352 mutation and 954 nsSNP sites could be mapped onto the structures. As compared to the average residues and nsSNPs, more disease mutations are found in protein cores (all residues: core/total = 24.5%; nsSNPs: core/total = 20.1%; disease mutations: core/total = 34.9%).

Disease-related mutations tend to be clustered (as measured by the clustering index described in the methods section), in contrast to nsSNPs, which are not. In 97 out of 667 (14%) structures in which at least 2 mutations can be mapped onto the structure, disease-related mutations are significantly (0.05% significance) clustered; in comparison, in only 4 out of 205 structures with at least 2 nsSNPs mapped, nsSNPs are clustered together at the same significance threshold. In experimentally determined structures, disease-related mutations are

significantly clustered in 27 out of 145 structures (19%), while nsSNPs are significantly clustered in only 2 out of 36 structures (6%). Both results show that no matter if predicted models were included for statistics or not, disease-related mutations are more significantly clustered together than nsSNPs.

As an example, Figure 2 shows the mapping of mutations and nsSNPs of Glutamate dehydrogenase 1 protein (SwissProt accession number P00367) onto its X-ray structure (PDB code 1l1f, chain A). This protein has 10 disease related mutations (associated with hyperinsulinism-hyperammonemia syndrome, HHS and highlighted red in Figure 2); all are closely located together, and the largest component of the mutation graph has 7 residues (with P-value = 0).

Figure 3 shows a model of a SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily A-like protein 1 (SwissProt accession number Q9NZC9**)** and the distribution of disease-related mutations and SNPs on this model. This protein has 10 mutations related to SIOD (Schimke immuno-osseous dysplasia) disease and 3 SNPs (collected in dbSNP) that can be mapped to the model. Analysis of the figure, supported by our statistical analysis, clearly shows that the disease mutations are clustered together (the largest component has 4 residues, P-value = 0.005), while the three SNPs do not (the largest component has 1 residue). Interestingly, one of the SNPs is located in the interface with many disease mutations; it suggests the possibility that this SNP may be deleterious as well. This example shows a possible way to study the effects of nsSNPs by comparing their spatial distribution with that of known disease mutation sites.

3.3. *Analysis of complexes*

The results from the modeling of human disease proteins and the analysis of the relative positions of nsSNPs and disease mutations on the structures strongly suggest that clusters of such mutations form specific patches on the surface and prompt the speculation that these patches are involved in protein-protein interactions. While a statistically rigorous evaluation of this hypothesis is not possible at present, we looked at details of a few available examples of experimentally determined complexes between human disease proteins.

We analyzed eight experimentally determined structures of complexes of disease proteins. This number seems rather small as compared to the number of potential protein-protein interactions collected in databases such as OPHID (http://ophid.utoronto.ca) [19] (its 5/2005 version collected 8836 protein-protein interactions involving at least one disease protein, and 1012 involving two disease proteins). This discrepancy illustrates the experimental difficulties involved in experimental studies of protein complex formations.

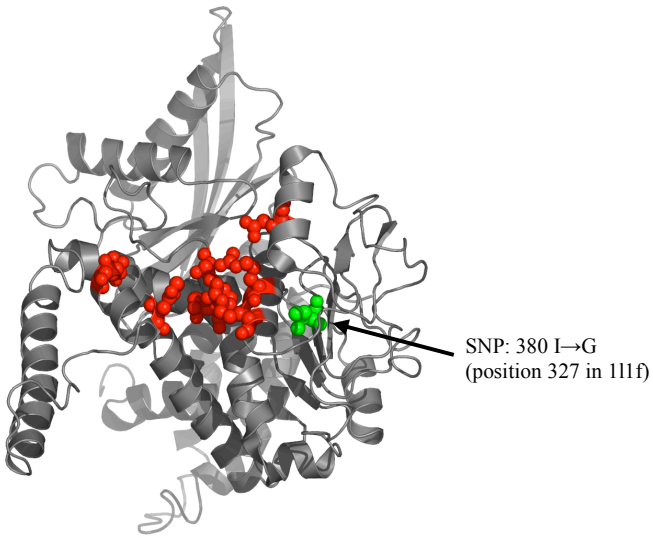

Figure 2. The mapping of mutations (in red ball-and-sticks) and nsSNPs (in green ball-and-stick) on the X-ray structure of SwissProt protein P00367 (PDB code 1l1fA)

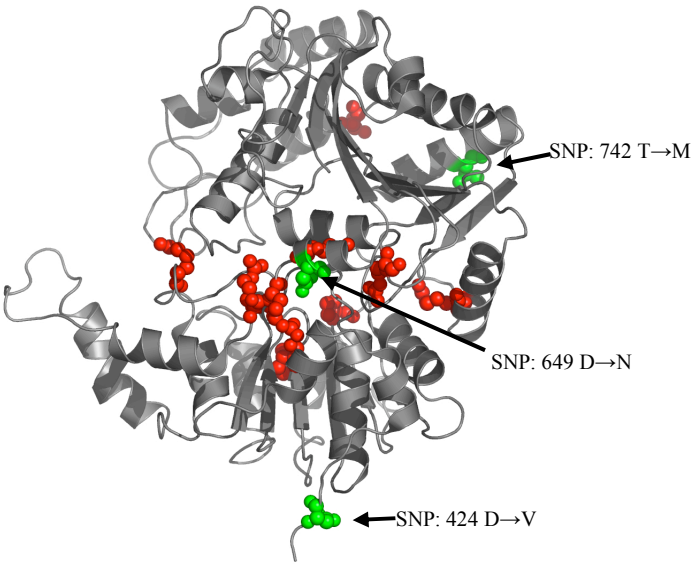

Figure 3. The mapping of mutations (in red ball-and-sticks) and nsSNPs (in green ball-and-stick) on the model of matrix-associated actin-dependent regulator of chromatin (SwissProt accession number Q9NZC9).

We found that most of such complexes show strong clustering of mutations around the binding interface. For instance, as shown in Figure 4, proteins integrin beta-3 (P05106) and integrin alpha-IIb (P08514) both have a lot of mutations associated with Glanzmann thrombasthenia (GT), the most common inherited disease of platelets, and these mutations (red ball-and-sticks) tend to be located in their binding interface; in contrast, three SNPs (green ball-and-sticks) are farther away from the binding interface.

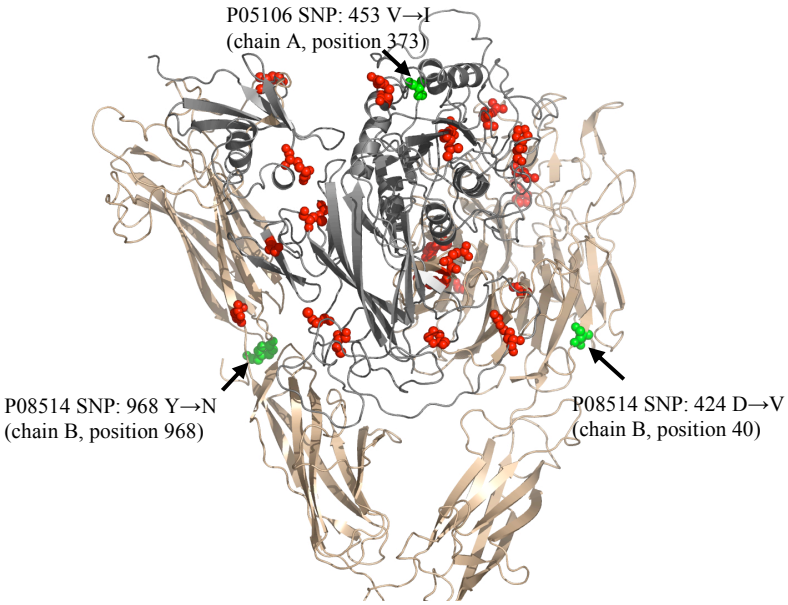

Figure 4. Clusters of disease mutations (red) on the structure of a complex (PDB code 1m1x) of two disease proteins, proteins: integrin beta-3 (P05106) and integrin alpha-IIb (P08514). Three SNPs highlighted in green ball-and-stick are also shown for comparison.

Another example is provided by the structure of a complex between GTPase-activating protein (P20936) and transforming protein p21/H-Ras-1 (P01112). Both proteins contain multiple domains and are implicated in a variety of human tumors. Their interaction is listed in the OPHID database, and more importantly, there is an X-ray structure of the complex between two domains one from each proteins (PDB code 1wq1). We mapped the mutations and SNPs of P20936 and P01112 onto structure 1wq1. Three mutations of P01112 are located on the domains present in the complex, forming two unique sites located exactly at the center of the interaction interface between protein P20936 and P01112 (see Figure 5). There are no experimentally determined structures for P09619

and for P20936 (outside the region of 718-1037). Domain analysis shows that P09619 has Igc2 domain and TyrKc domain (Tyrosine kinase domain), and P20936 has SH3 and SH2 domains (as well as several other domains and RasGAP domain) (see Figure 5). Homology searching shows that the SH3 and SH2 domains of P20936 and the TyrKc domain of P09619 can match the SH3, SH2 and TyrKc domains of protein Src Family Kinase (with known X-ray structure, PDB code 2hck), respectively. It is known that in 2hck, the so-called tail peptide (highlighted in blue in Figure 5) of the catalytic domain (TyrKc domain) interacts with the binding sites (highlighted in red) in SH2 domain of the same protein, helping to lock the structure in an inactive form (autoinhibition) [20,21]. Interestingly, three out of four remaining disease-related SNPs in P20936 are matched to the tail peptide binding sites in the SH2 domain of 2hck, suggesting that these three disease related SNPs are located in the binding interface of P20936 and P09619 (involving the tail peptide). We mapped the mutations onto structure 2hck (which may be used as a template for modeling the complex of P20936 and P09619) in Figure 5, considering it will be difficult to precisely model the interaction of the short tail peptide with its binding site. In summary, the disease mutations of P20936 and P01112 are mainly located in the binding interface of the interaction network composed of these two proteins and P09619.

4. Conclusion

We have generated three-dimensional models for over a thousand human disease proteins and set up a publicly available Web site, showing annotated pictures for all the models. The current structural coverage of the human disease proteins is close to 50% of the total residues. We expect that the coverage would increase with the continuous growth of structural databases.

Our analysis of the spatial distribution of disease mutations shows their non-uniform distribution, and in particular forming patches on surfaces of proteins. It is tempting to speculate that such patches are located at or near protein-protein interaction interfaces.

To test this hypothesis we evaluated a number of structures of complexes, initially focusing on experimentally determined structures. The number of such complexes is rather small, but the examples are very suggestive. Indeed, in most cases disease mutations cluster at binding interfaces.

Recently, tremendous advances have been achieved in identifying the protein-protein interactions, both from large scale experiments and computational approach [22-24]. Thousands of protein-protein interactions involving disease proteins have been identified or predicted (as collected in

OPHID database). The huge gap between the number of potential interactions and the number of experimentally determined structures of such complexes suggests that modeling would play an important role in filling the gap. One possibility is to use the structure of a complex as a template [25,26] (as suggested by the example shown in Figure 5). But this method is limited because only a relatively small number of complexes are available for modeling. Large scale *ab initio* modeling of protein complexes would be necessary to further evaluate our hypothesis. We plan to extend the current study by using *ab initio* docking methods, such as GRAMM [27], to predict the structures of complexes of disease proteins. Another possibility is to use the non-uniform distribution of disease mutations and SNPs in this process as an additional guidance. Also we will search for other alternative ways of building models for protein-protein interactions [28], such as fitting models to the low resolution complex structure from electron microscopy (EM) when data is available. The resulting models of complexes would become an important resource for studying the functions of disease proteins and the mechanism of diseases.

Acknowledgments

We thank Dr. Lukasz Jaroszewski for his help with the protein modeling. This project was supported by NIH grant P01 GM63208.

References

1. M. Cargill, D. Altshuler, J. Ireland, P. Sklar, K. Ardlie, N. Patil, N. Shaw, C.R. Lane, E.P. Lim, N. Kalyanaraman, J. Nemesh, L. Ziaugra, L. Friedland, A. Rolfe, J. Warrington, R. Lipshutz, G.Q. Daley, and E.S. Lander, *Nature Genetics*. **22**(3), 231-238 (1999)
2. A.C. Martin, A.M. Facchiano, A.L. Cuff, T. Hernandez-Boussard, M. Olivier, P. Hainaut, and J.M. Thornton, *Hum Mutat*. **19**(2), 149-164 (2002)
3. M.A. Fleming, J.D. Potter, C.J. Ramirez, G.K. Ostrander, and E.A. Ostrander, *Proc Natl Acad Sci U S A*. **100**(3), 1151-1156 (2003)
4. P.C. Ng and S. Henikoff, *Nucleic Acids Res*. **31**(13), 3812-3814 (2003)
5. S. Sunyaev, V. Ramensky, and P. Bork, *Trends Genet*. **16**(5), 198-200 (2000)
6. Z. Wang and J. Moult, *Hum Mutat*. **17**(4), 263-270 (2001)
7. A. Cavallo and A.C. Martin, *Bioinformatics*. **21**(8), 1443-1450 (2005)
8. S.D. Mooney and R.B. Altman, *Bioinformatics*. **19**(14), 1858-1860 (2003)
9. R. Karchin, M. Diekhans, L. Kelly, D.J. Thomas, U. Pieper, N. Eswar, D. Haussler, and A. Sali, *Bioinformatics*. **21**(12), 2814-2820 (2005)
10. Y.L. Yip, H. Scheib, A.V. Diemand, A. Gattiker, L.M. Famiglietti, E. Gasteiger, and A. Bairoch, *Hum Mutat*. **23**(5), 464-470 (2004)

11. L. Rychlewski, L. Jaroszewski, W. Li, and A. Godzik, *Protein Science*. **9**, 232-241 (2000)
12. Z. Xiang and B. Honig, *J Mol Biol*. **311**(2), 421-430 (2001)
13. A. Sali and T.L. Blundell, *J Mol Biol*. **234**, 779-815 (1993)
14. L. Jaroszewski, K. Pawlowski, and A. Godzik, *J Mol Model*. **4**, 294 - 309 (1998)
15. R. Schwarzenbacher, A. Godzik, S.K. Grzechnik, and L. Jaroszewski, *Acta Crystallogr D Biol Crystallogr*. **60**(Pt 7), 1229-1236 (2004)
16. B. Lee and F.M. Richards, *J Mol Biol*. **55**, 379-400 (1971)
17. E. Minieka, *Optimization algorithms for networks and graphs*. New York: Marcel Dekker (1978)
18. I. Friedberg, L. Jaroszewski, Y. Ye, and A. Godzik, *Curr Opin Struct Biol*. **14**(3), 307-312 (2004)
19. K.R. Brown and I. Jurisica, *Bioinformatics*. **21**(9), 2076-2082 (2005)
20. T. Pawson, *Nature*. **385**(6617), 582-583 (1997)
21. W. Xu, S.C. Harrison, and M.J. Eck, *Nature*. **385**(6617), 595-602 (1997)
22. P. Uetz, L. Giot, G. Cagney, T.A. Mansfield, R.S. Judson, J.R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. Yang, M. Johnston, S. Fields, and J.M. Rothberg, *Nature*. **403**(6770), 623-627 (2000)
23. C. von Mering, R. Krause, B. Snel, M. Cornell, S.G. Oliver, S. Fields, and P. Bork, *Nature*. **417**(6887), 399-403 (2002)
24. H. Yu, N.M. Luscombe, H.X. Lu, X. Zhu, Y. Xia, J.D. Han, N. Bertin, S. Chung, M. Vidal, and M. Gerstein, *Genome Res*. **14**(6), 1107-1118 (2004)
25. P. Aloy and R.B. Russell, *Proc Natl Acad Sci U S A*. **99**(9), 5896-5901 (2002)
26. L. Lu, A.K. Arakaki, H. Lu, and J. Skolnick, *Genome Res*. **16**(6A), 1146-1354 (2003)
27. E. Katchalski-Katzir, I. Shariv, M. Eisenstein, A.A. Friesem, C. Aflalo, and I.A. Vakser, *Proc Natl Acad Sci U S A*. **89**(6), 2195-2199 (1992)
28. R.B. Russell, F. Alber, P. Aloy, F.P. Davis, D. Korkin, M. Pichaud, M. Topf, and A. Sali, *Curr Opin Struct Biol*. **14**(3), 313-324 (2004)

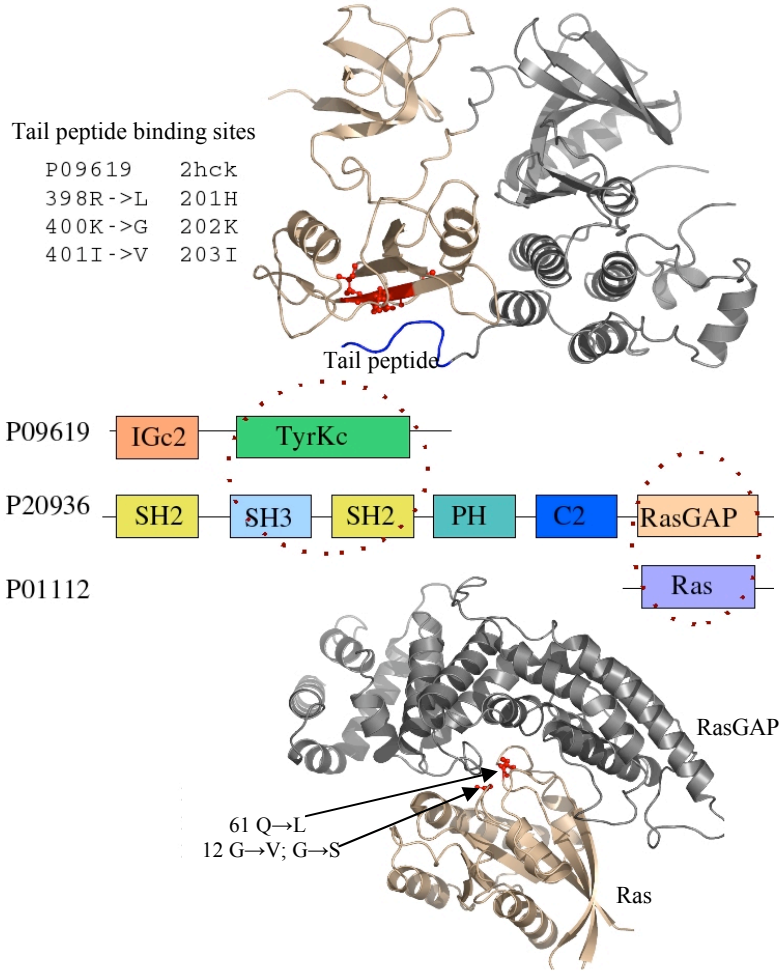

Figure 5. The interaction network of GTPase-activating protein (P20936), transforming protein p21/H-Ras-1 (P01112) and alpha-hemolysin **(**P09619). Two complexes are involved: one is the complex between RasGAP domain of protein P20936 and Ras domain of protein P01112 (PDB code 1w1q), with disease-related mutations highlighted in red ball-and-stick in the graph (shown in the bottom of this figure), and the other one is the complex between SH3 and SH2 domains of protein P20936 and the TyrKc domain of P09619 (PDB code 2hck, shown in the top of this figure).