Integration of genome data and protein structures: prediction of protein folds, protein interactions and 'molecular phenotypes' of single nucleotide polymorphisms

Shamil Sunyaev*†‡§, Warren Lathe III*†# and Peer Bork*†¶

With the massive amount of sequence and structural data being produced, new avenues emerge for exploiting the information therein for applications in several fields. Fold distributions can be mapped onto entire genomes to learn about the nature of the protein universe and many of the interactions between proteins can now be predicted solely on the basis of the genomic context of their genes. Furthermore, by utilising the new incoming data on single nucleotide polymorphisms by mapping them onto three-dimensional structures of proteins, problems concerning population, medical and evolutionary genetics can be addressed.

Addresses
*European Molecular Biology Laboratory (EMBL), Meyerhofstrasse 1, 69117 Heidelberg, Germany
†Max-Delbrueck Centre for Molecular Medicine (MDC), Robert-Roessle-Strasse 10, 13122 Berlin, Germany
‡Engelhardt Institute of Molecular Biology (IMB), Vavilova 32, 117984 Moscow, Russia
§e-mail: sunyaev@embl-heidelberg.de
#e-mail: lathe@embl-heidelberg.de
¶e-mail: bork@embl-heidelberg.de

Current Opinion in Structural Biology 2001, 11:125–130

0959-440X/01/$ – see front matter
Published by Elsevier Science Ltd.

Abbreviations

cSNP	coding SNP
PDB	Protein Data Bank
SNP	single nucleotide polymorphism

Introduction

The past several years have been marked by the successful completion of numerous genome projects, ranging from the short genomes of prokaryotes to our own. As a result of this extraordinary growth, different types of genomic data (Figure 1), including sequences of complete genomes, complete sets of proteins (proteomes) and data on genetic variation, have become available for analysis. The simultaneous growth of structural data also provides the possibility of performing this analysis from a structural perspective. In this review, we focus on the impact of genomic data on studies of protein structures and interactions and, perhaps more importantly, on the new applications of the structural research and the challenges raised by these applications. Currently, three trends bridge the structural and genomics fields: analysis of protein folds in complete genomes; use of genome information for predicting protein–protein interactions; and structural analyses of disease mutations and single nucleotide polymorphisms (SNPs). We only briefly introduce the first two topics, because they have been extensively covered in many reviews, and mostly focus on the issue of integrating structural analysis with research on human genetic variation.

Predicting the number of protein folds in genomes using homology searches

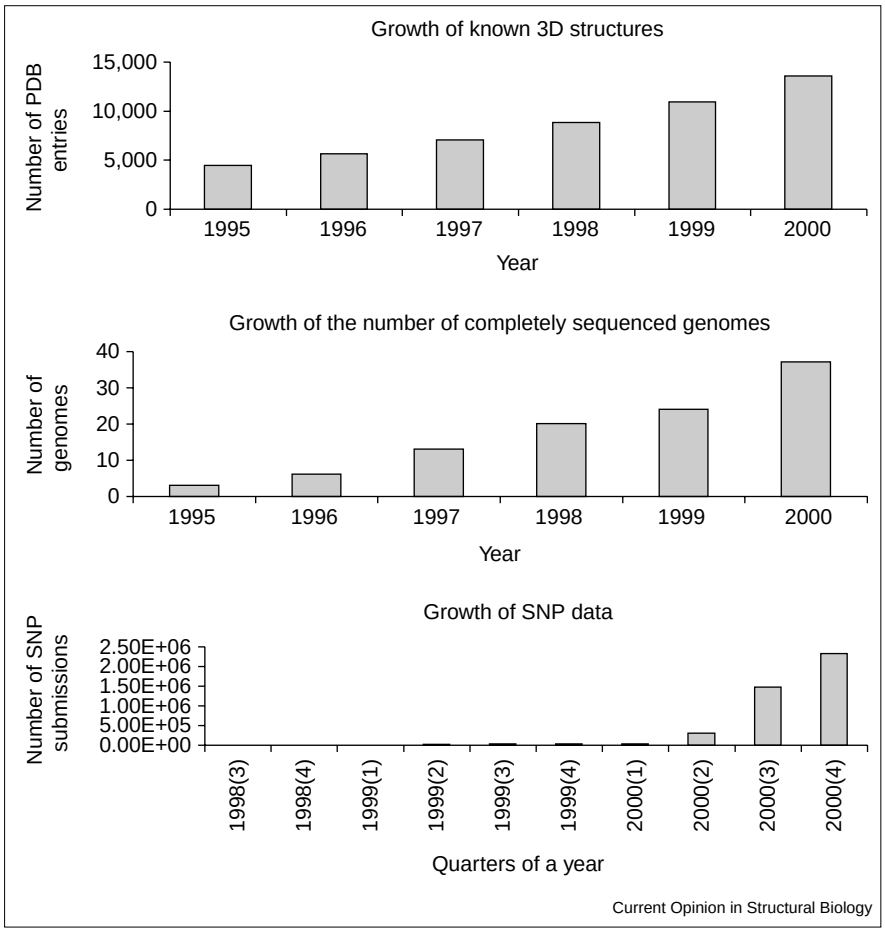
Soon after the first complete genome sequences became publicly available, it was realised that they represent datasets for the analysis of protein folds that are statistically far more natural than compared to the PDB [1]. Furthermore, iterative database search methods such as PSI-BLAST [2] had been developed and a number of reports revealed that the folds of more than 30% of the proteins in prokaryotic genomes can be reliably predicted using homology-based techniques [3–6]. These iterative methods are at least twice as sensitive as pairwise protein comparisons [7]. This increased sensitivity can sometimes identify the common evolutionary origin of proteins that were previously thought to belong to different superfamilies and whose structural similarities were assumed to be a result of parallel evolution (analogous rather than homologous structures). An example of a newly identified nontrivial homology relationship involves enzymes of the TIM-barrel fold from the central metabolism that have been shown to share a common evolutionary origin despite being grouped into 12 distinct SCOP superfamilies [8,9].

Fold predictions for complete genomes allowed a new estimate of the total number of protein folds and of the number of protein folds in individual genomes [10]. The estimate qualitatively agrees with that of earlier studies [11,12]. The statistical analysis [13] shows that the nature of the fold distribution is universal for all genomes and agrees with earlier theoretical considerations [14]. Although these estimates give a much clearer picture of the nonuniform fold distribution and the low number of folds, it remains to be shown whether the physical constraints of the polypeptide chain or the specific features of protein evolution led to the current fold repertoire.

Predicting protein interactions using genomic context

The availability of numerous complete genome sequences also enables comparative analysis to predict various functional features at the protein level. The genomic context of genes reveals the physical, functional and genetic interactions of the respective gene products. Several strategies have been used to explore genomic context. First, gene fusion in one species is indicative of a protein–protein interaction between the gene products of the two fused genes and can be

Figure 1



The growth of structural data (counted in PDB entries) in recent years is compared with the growth of genomic data, represented by the number of completely sequenced genomes and by the number of known human SNPs. Although all types of presented data accumulate fast, the growth of genomic data, especially SNP data, outperforms structural data accumulation. Data kindly provided by A Brookes (Karolinska Institute), S Sherry (NCBI) and M Huynen (EMBL).

assumed for all the orthologues of the two genes in other species [15,16]. Second, conservation of gene neighbourhood in some divergent species also strongly indicates the occurrence of an interaction between the two encoded gene products, even if they are not neighbours in many other genomes [17,18]. In fact, entire subcellular systems can be identified if neighbourhood information for different genes is systematically merged [19]. Third, the co-occurrence (this has also been coined phylogenetic profile or COG pattern) of two genes (and their orthologues) in the same subset of species indicates that the two genes interact [20–22]. Fourth, the presence of shared regulatory elements hints at the co-regulation of the respective downstream genes and, hence, at the genetic or functional interaction of the respective gene products [23].

All these strategies and the respective methods can be combined to increase their sensitivity [24]. Although context-based methods are not yet as powerful as classical homology-based function prediction methods [25] and differences between prokaryotic and eukaryotic evolution have to be considered, the power of genomic context analysis will increase with each new genome published.

Structural analysis of allelic variants

A very important part of genomic research, which was underestimated at the beginning of the genomics era, is the analysis of genetic variation in populations. The studies of genetic variation are now appreciated in the context of the human genome project and several consortia around the world have already identified more than two million DNA variants in the human population (Figure 1). Therefore, it is time to integrate these data with information from other resources, such as three-dimensional structures of proteins. The application of structural data to research on genetic variation can propel studies on the identification of the genetic roots of phenotypic variation and bring new insights to the fields of population and evolutionary genetics. In turn, structural research may also benefit from using mutation data. Catalogues of naturally occurring mutations with known phenotype association might, in some cases, substitute for experiments on site-directed mutagenesis in studies of protein folding and binding.

In the following, we give a brief introduction to the nature of mutation and polymorphism data, and then review the first approaches to combine them with structural

information in both specific case studies and general analyses of amino-acid-replacing SNPs mapped to 3D structures of proteins.

Mutations

For a long time, data on genetic variation were available only for particular loci mainly corresponding to genes associated with simple monogenic diseases. Numerous locus-specific databases contain mutation data, often from patients with genetic disorders. These databases report not only the various mutations and polymorphisms identified in the locus, but also the phenotypic features associated with these genetic variants. Locus-specific databases can be easily accessed through the Human Gene Mutation Database (http://archive.uwcm.ac.uk/uwcm/mg/hgmd0.html) [26] or through SRS (http://srs.ebi.ac.uk/) [27].

The Online Mendelian Inheritance in Man (OMIM) database (http://www3.ncbi.nlm.nih.gov/omim) [28] provides information on allelic variants, together with phenotypic or functional associations if available. This information is extracted from the literature and covers a large number of human genes.

Locus-specific mutation databases now contain thousands of amino acid variants in well-characterised proteins. On one hand, structural analysis of these data can help determine residues that are crucial for specific interactions or for the formation of native protein structure. On the other hand, structural and functional analysis of amino acid substitutions collected in locus-specific databases can reveal the molecular background of particular genetic diseases, determine the major mechanisms responsible for the destructive effect of disease-causing mutations and, possibly, help develop prediction algorithms.

Polymorphism

The massive amount of data on human DNA variants that have become available as a result of large genome projects require the problem to be approached from a different direction. These data mostly do not contain any information on phenotypes. Unlike locus-specific databases, however, they provide a true large-scale picture of human genetic variation mainly represented by SNPs. Novel SNP-centred databases, such as the Human Genic Bi-allelic Sequences Database (HGBASE) (http://hgbase.cgr.ki.se) [29••] and dbSNP (http://www.ncbi.nlm.nih.gov/SNP/) [30••], have already accumulated more than two million allelic variants, most of them identified very recently, and many more are expected to come soon (Figure 1).

SNPs are DNA allelic variants that arise as a result of single nucleotide mismatches and that have an appreciable allele frequency in the population [31]. There are several reasons why SNPs are the focus of many studies on human genetics. First, only a minor fraction of human disorders are simple monogenic diseases. The majority of human disease phenotypes are now believed to be of a complex nature, involving common DNA variants together with environmental factors. The identification of variants that increase susceptibility to human diseases is one of the key problems in medical genetics. Second, analysis of genetic variation in a population can help in the pursuit of solutions to many problems in evolutionary genetics. Third, knowledge of genetic variation in the modern human population is a clue to our understanding of human origins and features of the human population in prehistoric times.

The large amounts of data on SNPs create new challenges for research on structures and interactions in probably the most important problem of DNA variation studies, namely linking genetic variation with phenotypic variation and then with natural selection. Obviously, the phenotypic effect of a nucleotide substitution is always caused by structural or functional changes in DNA, RNA or protein. Although nucleotide substitutions in many regions of noncoding DNA (especially in many untranslated regions) can be functionally important, analysis of the allele frequency spectrum suggests that selectively non-neutral alleles present in the population are much more frequent among protein allelic variants. Only a fraction of SNPs fall in this category, namely the subfraction of those SNPs that occur in the coding sequence (coding SNPs, cSNPs) and that lead to amino acid replacements (nonsynonymous cSNPs). Thus, analysis of 'molecular phenotypes', that is, allele-specific features in structure, folding, binding or stability, can help to explain the biological mechanism of phenotypic effect or even to predict this effect (these predictions can be used for prioritisation of candidates for epidemiological association studies).

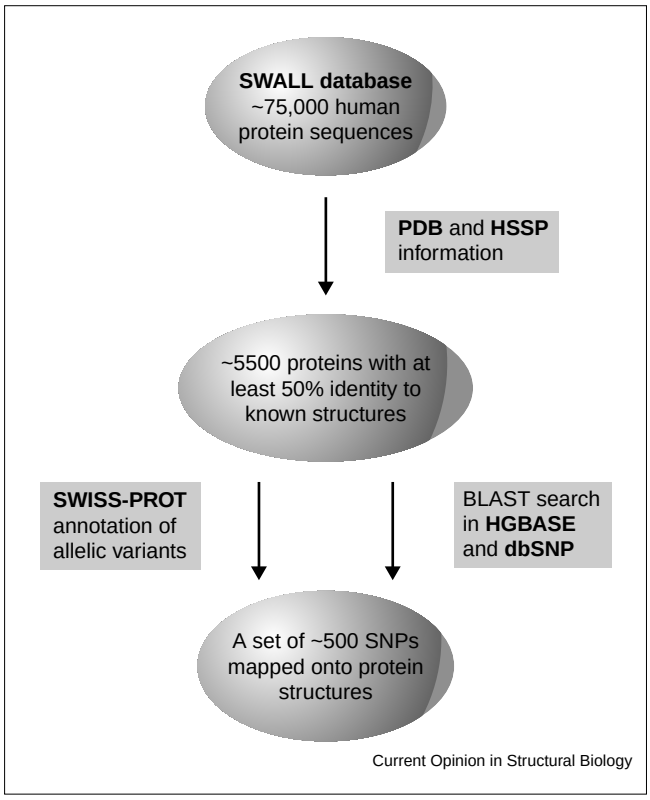
Case studies

Several recent case studies have shown that a detailed analysis of the effect of SNPs on protein function can be fruitful in gaining an understanding of the causes of human disease.

The functional importance of an SNP can also be a result of an effect on specific sites in noncoding DNA (e.g. regulatory sites) or RNA structures. Shen *et al.* [32] recently provided evidence that some SNPs can cause different structural folds of mRNA. One good example of such a disease-causing SNP in a noncoding region is the tau gene. The structure of tau exon 10 splicing regulatory element RNA has recently been deciphered and has been shown to form a stable folded stem-loop structure [33•]. Several mutations occur in the noncoding intron after exon 10 and lead to dementing diseases. This work showed that these mutations destabilise the described RNA stem-loop structure and this destabilisation increases the splicing in of exon 10.

A number of recent case studies detailing the effect of single nucleotide mutations on the structure and function

Figure 2



Mapping nonsynonymous cSNPs (amino acid allelic variants) onto 3D structures of proteins. Three sources of information about allelic variants have been used: SNP databases HGBASE and dbSNP, and the annotation of allelic variants in SWISS-PROT (only clear polymorphisms were considered, whereas disease mutations were excluded from the analysis). Numbers correspond to June 2000 and, as a result of the extraordinary data growth, the number of SNPs that can be mapped onto 3D structures is likely to increase considerably. However, direct projection is impossible because the majority of newly identified SNPs are randomly spaced in the genome, so only a minor fraction of them are nonsynonymous cSNPs.

of proteins have not only shown the specific structural causes of disease, but have also given insights into the mechanisms of the native proteins. For example, Gooptu *et al.* [34•] have described the structural and biochemical characteristics of a point mutation (Leu55Pro) in α_1-antichymotrypsin, a protease inhibitor of the serpin superfamily, that explain the loss of activity of plasma α_1-antichymotrypsin and the chronic obstructive pulmonary disease in patients with this mutation. Another of several examples includes a single point mutation in human apolipoprotein A-I (apoA-I) that is associated with coronary artery disease. Recent work [35,36] has pinpointed the structural and functional effects of this mutation.

These and other recent studies on the effects of single point mutations on protein structure and function have given us insights into both the cause of disease and the functions of proteins [37•,38–42].

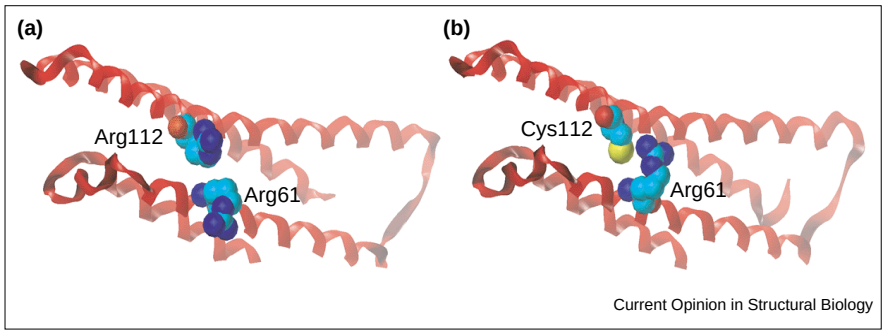
Systematic structural analysis of single nucleotide polymorphisms

As the preceding case studies might indicate, the increase in structural and genomic data allows a more systematic approach to this area of research (Figure 2). Systematic attempts to map nonsynonymous cSNPs (amino-acid-replacing SNPs) onto 3D structures of proteins and to analyse the possible impact of allelic variants on protein structure or function have recently been made. Sunyaev *et al.* [43••] mapped a set of nonsynonymous cSNPs from public databases onto 3D structures of the corresponding proteins and analysed their structural location, together with sequence conservation of the SNP sites in homologous protein families. Structural and conservation characteristics of SNP sites have been then compared to characteristics of amino acid substitutions between human proteins and their closely related orthologues (similar to the McDonald and Kreitman test [44] used in population genetics to detect selection) and to characteristics of mutations known to be responsible for diseases. It was shown that the number of SNPs located in structurally or functionally important sites (buried sites, conservative sites, etc.) is significantly higher compared with between-species substitutions (although it is obviously lower compared with disease mutations). Results of this analysis suggest that a significant fraction of nonsynonymous cSNPs are likely to affect protein structure or function, and thus might constitute alleles deleterious for phenotype. Later research by the same group [45••] shows that deleterious amino acid variants can be usefully predicted from simple structural considerations, together with conventional sequence analysis techniques. The study resulted in the estimate of approximately 2000 deleterious allelic variants per genome of an average individual, with an average for these variants selection coefficient against heterozygotes of approximately 10^{-3}.

Wang and Moult [46••] studied structural features of a number of nonsynonymous cSNPs and compared them with a set of disease-causing mutations. They have developed a rule-based model to identify amino acid substitutions with a negative effect on the structure and function of the protein, and classified allelic variants into neutral and deleterious; deleterious variants were subsequently classified according to the effect on stability, binding, catalysis, allosteric response or post-translational modification. Surprisingly, the most deleterious amino acid variants were shown to affect the stability of the protein. The specific effect of these variants on protein structure has been further analysed.

These are the first large-scale applications of structural biology to the analysis of genetic variation. The future potential of systematic structural studies of SNPs for both medical and population genetics is based on the importance of the functional and phenotypic characterisation of DNA variants.

Figure 3



A human polymorphism that is likely to represent an advantageous compensatory substitution is shown as an example of the evolutionary implication of structural data. Representations of two alleles of human apolipoprotein E (ApoE), (a) ApoE4 and (b) ApoE3, are displayed. ApoE4 is likely to be an ancestral allele (as suggested by the sequences of great apes) [47,48]. This allele is associated with the elevated risk of Alzheimer's disease and was also hypothesised to affect reproductive efficiency [47]. All known animal sequences, including that of the apes, suggest a recent human-specific substitution for arginine in codon 61, a position critical for domain interaction [47]. The ApoE3 allele has a substitution of Arg112 for Cys112. As a result of selective advantage, this probably younger allele achieved higher frequency in the present-day population. The 3D structure reveals that the Arg112Cys substitution has a compensatory nature. It removes the close and unfavourable placement of two arginine residues at this site. Therefore, the substitution for Cys112 compensates the deleterious effect of Arg61.

Owing to the expected huge numbers of SNPs, it is technically impossible to proceed with association studies for all SNPs found in the human population in order to relate them to phenotypic features. The same holds for the large-scale experimental functional analysis of all genetic variants. In contrast, given the large amount of protein 3D structures to be determined in the near future, computational structural analysis of large numbers of allelic variants seems feasible. This can help the study of genotype/phenotype relationships through the analysis of 'molecular phenotypes'. Analysis of protein–protein interactions can, in turn, shed light on the problem of multigenic phenotypes (e.g. complex human disorders).

The amount of genetic variation that is subject to natural selection is a long-standing problem in population genetics. First attempts to predict deleterious alleles and to estimate their impact on fitness have been described above. In addition to the prediction and analysis of deleterious alleles, structural studies on allelic variants can probably help to understand, in some cases, the molecular basis of positive Darwinian selection (an example of compensatory advantageous substitution is shown in Figure 3) and balancing selection. Thus, the integration of SNPs into structural analysis can have implications not only for molecular medicine, but also for protein evolution.

Conclusions

The current bottleneck for structural studies using genomic data, that is, the deficit in known protein structures, will be widened with future structural genomics projects. The studies described here can be considered as pilot studies that show the importance and utility of a systematic structural genomics initiative beyond just the reason of structure-based function predictions. The selection of targets for structural genomics should include proteins with as yet unpredicted folds, proteins involved in important interactions and pathways or, for example, proteins with allelic variants likely to be associated with phenotypes.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Gerstein M: A structural census of genomes: comparing bacterial, eukaryotic, and archaeal genomes in terms of protein structure. *J Mol Biol* 1997, 274:562-576.
2. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997, 25:3389-3402.
3. Huynen M, Doerks T, Eisenhaber F, Orengo C, Sunyaev S, Yuan Y, Bork P: Homology-based fold predictions for ***Mycoplasma genitalium*** proteins. *J Mol Biol* 1998, 280:323-326.
4. Teichmann SA, Park J, Chothia C: Structural assignments to the ***Mycoplasma genitalium*** proteins show extensive gene duplications and domain rearrangements. *Proc Natl Acad Sci USA* 1998, 95:14658-14663.
5. Rychlewski L, Zhang B, Godzik A: Fold and function predictions for ***Mycoplasma genitalium*** proteins. *Fold Des* 1998, 3:229-238.
6. Wolf YI, Brenner SE, Bash PA, Koonin EV: Distribution of protein folds in the three superkingdoms of life. *Genome Res* 1999, 9:17-26.
7. Park J, Karplus K, Barrett C, Hughey R, Haussler D, Hubbard T, Chothia C: Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J Mol Biol* 1998, 284:1201-1210.
8. Lo Conte L, Ailey B, Hubbard TJ, Brenner SE, Murzin AG, Chothia C: SCOP: a structural classification of proteins database. *Nucleic Acids Res* 2000, 28:257-259.
9. Copley RR, Bork P: Homology among (betaalpha)(8) barrels: implications for the evolution of metabolic pathways. *J Mol Biol* 2000, 303:627-641.
10. Wolf YI, Grishin NV, Koonin EV: Estimating the number of protein folds and families from complete genome data. *J Mol Biol* 2000, 299:897-905.
11. Chothia C: Proteins. One thousand families for the molecular biologist. *Nature* 1992, 357:543-544.
12. Orengo CA, Jones DT, Thornton JM: Protein superfamilies and domain superfolds. *Nature* 1994, 372:631-634.
13. Koonin EV, Wolf YI, Aravind L: Protein fold recognition using sequence profiles and its application in structural genomics. *Adv Protein Chem* 2000, 54:245-275.
14. Govindarajan S, Recabarren R, Goldstein RA: Estimating the total number of protein folds. *Proteins* 1999, 35:408-414.

15. Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D: Detecting protein function and protein-protein interactions from genome sequences. *Science* 1999, 285:751-753.

16. Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA: Protein interaction maps for complete genomes based on gene fusion events. *Nature* 1999, 402:86-90.

17. Dandekar T, Snel B, Huynen M, Bork P: Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci* 1998, 23:324-328.

18. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N: The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci USA* 1999, 96:2896-2901.

19. Lathe WC, Snel B, Bork P: Gene context conservation of a higher order than operons. *Trends Biochem Sci* 2000, 25:474-479.

20. Tatusov RL, Koonin EV, Lipman DJ: A genomic perspective on protein families. *Science* 1997, 278:631-637.

21. Huynen MA, Bork P: Measuring genome evolution. *Proc Natl Acad Sci USA* 1998, 95:5849-5856.

22. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO: Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci USA* 1999, 96:4285-4288.

23. McGuire AM, Hughes JD, Church GM: Conservation of DNA regulatory motifs and discovery of new motifs. *Genome Res* 2000, 10:744-757.

24. Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D: A combined algorithm for genome-wide prediction of protein function. *Nature* 1999, 402:83-86.

25. Huynen M, Snel B, Lathe W III, Bork P: Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res* 2000, 10:1204-1210.

26. Krawczak M, Cooper DN: The human gene mutation database. *Trends Genet* 1997, 13:121-122.

27. Lehvaslaiho H, Ashburner M, Etzold T: Unified access to mutation databases. *Trends Genet* 1998, 14:205-206.

28. Online Mendelian Inheritance in Man (OMIM™) on World Wide Web URL: http://www3.ncbi.nlm.nih.gov/omim/

29. •• Brookes AJ, Lehvaslaiho H, Siegfried M, Boehm JG, Yuan YP, Sarkar CM, Bork P, Ortigao F: HGBASE: a database of SNPs and other variations in and around human genes. *Nucleic Acids Res* 2000, 28:356-360.
HGBASE is a database that accumulates human polymorphism data from both submissions and literature. The records are highly curated and annotated.

30. •• Smigielski EM, Sirotkin K, Ward M, Sherry ST: dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Res* 2000, 28:352-355.
The authors discuss the dbSNP database – currently the most comprehensive source of SNP data.

31. Brookes AJ: The essence of SNPs. *Gene* 1999, 234:177-186.

32. Shen LX, Basilion JP, Stanton VP Jr: Single-nucleotide polymorphisms can cause different structural folds of mRNA. *Proc Natl Acad Sci USA* 1999, 96:7871-7876.

33. • Varani L, Hasegawa M, Spillantini MG, Smith MJ, Murrell JR, Ghetti B, Klug A, Goedert M, Varani B: Structure of tau exon 10 splicing regulatory element RNA and destablization by mutation of frontotemporal dementia and parkinsonism linked to chromosome 17. *Proc Natl Acad Sci USA* 1999, 96:8229-8234.
The authors report the case of a noncoding single nucleotide polymorphism that has an effect on phenotype through the change in the splicing.

34. • Gooptu B, Hazes B, Chang WW, Dafforn TR, Carrell RW, Read RJ, Lomas DA: Inactive conformation of the serpin indicates two-stage insertion of the reactive loop: implications for inhibitory function and conformational disease. *Proc Natl Acad Sci USA* 2000, 97:67-72.
Structural and biochemical analysis of a point mutation helps to explain the loss of activity of α_1-antichymotrypsin. This provides the structural basis of the chronic obstructive pulmonary disease.

35. Cho KH, Jonas A: A key point mutation (V156E) affects the structure and functions of human apolipoprotein A-I. *J Biol Chem* 2000, 275:26821-26827.

36. Dong LM, Weisgraber KH: Human apolipoprotein E4 domain interaction. Arginine 61 and glutamic acid 255 interact to direct the preference for very low density lipoproteins. *J Biol Chem* 1996, 27:19053-19057.

37. • Stebbins CE, Kaelin WG, Pavletich NP: Structure of the VHL-ElonginC-ElonginB complex: implications for VHL tumor suppressor function. *Science* 1999, 284:455-461.
The structure of the VHL–ElonginC–ElonginB ternary complex is described. Mutation 'hot spots' in VHL have revealed important structural features of individual proteins and their interfaces, and have given insights into the functions and mechanism of VHL and the VHL–ElonginC–ElonginB complex. In particular, the tumor-associated Tyr98 hot spot, the second most frequently mutated residue in VHL, occurs on the surface of the protein at the opposite end from the ElonginC-binding site and has apparently no structural role. This and other observations suggest that this segment of residues corresponds to another binding site of VHL and is important for tumor suppression.

38. Takiguchi K, Itoh K, Shimmoto M, Ozand PT, Doi H, Sakuraba H: Structural and functional study of K453E mutant protective protein/cathepsin A causing the late infantile form of galactosialidosis. *J Hum Genet* 2000, 45:200-206.

39. Basmaciogullari S, Autiero M, Culerrier R, Mani J, Gaubin M, Mishal Z, Guardiola J, Granier C, Piatier-Tonneau D: Mapping the CD4 binding domain of gp17, a glycoprotein secreted from seminal vesicles and breast carcinomas. *Biochemistry* 2000, 39:5332-5340.

40. Musco G, Stier G, Kolmerer B, Adinolfi S, Martin S, Frenkiel T, Gibson T, Pastore A: Towards a structural understanding of Friedreich's ataxia: the solution structure of frataxin structure. *Fold Des* 2000, 8:695-707.

41. Di Barletta M, Ricci E, Galluzzi G, Tonali P, Mora M, Morandi L, Romorini A, Voit T, Orstavik KH, Merlini L *et al.*: Different mutations in the LMNA gene cause autosomal dominant and autosomal recessive Emery-Dreifuss muscular dystrophy. *Am J Hum Genet* 2000, 66:1407-1412.

42. Powell B, Soong R, Iacopetta B, Seshadri R, Smith DR: Prognostic significance of mutations to different structural and functional regions of the p53 gene in breast cancer. *Clin Cancer Res* 2000, 6:443-451.

43. •• Sunyaev S, Ramensky V, Bork P: Towards a structural basis of human non-synonymous single nucleotide polymorphisms. *Trends Genet* 2000, 16:198-200.
A set of amino-acid-replacing SNPs has been mapped onto 3D structures of proteins. The structural location and sequence conservation of SNP sites was compared with divergence data and disease mutations. The results suggest that a significant fraction of amino acid allelic variants affect the structure or function of proteins.

44. McDonald JH, Kreitman M: Adaptive protein evolution at the Adh locus in ***Drosophila***. *Nature* 1991, 351:652-654.

45. •• Sunyaev S, Ramensky V, Lathe WC III, Koch I, Kondrashov AS, Bork P: Prediction of human deleterious alleles. *Human Mol Genet* 2001, in press.
A method to predict the deleterious effect of amino acid allelic variants has been developed and applied to estimate both the number of deleterious variants in the average human genome and their mean selection coefficient.

46. •• Wang Z, Moult J: SNPs, protein structure, and disease. *Human Mutat* 2001, in press.
A rule-based model helped to identify amino acid allelic variants with an effect on protein structure/function/interactions. Statistics for different damaging effects are presented.

47. Mahley RW, Rall SC Jr: Apolipoprotein E: far more than a lipid transport protein. *Ann Rev Genom Human Genet* 2000, 1:507-538.

48. Rogaev EI, Korovaitseva GI, Sherbatych T, Dvoryanchikov G, Ryazanskaya N, Brusov O, Balaban P, Tyrsin O, Grigorenko A, Grivennikov IA *et al.*: Evolutionary and molecular-genetics analysis of genes for dementia of Alzheimer's type. *Howard Hughes Medical Institute Meeting of International Scholars*: *1999 June 22–26; Moscow.* Abstract P.83.