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
assumed for all the orthologues of the two genes in other species [15,16]. Second, conservation of gene neighbour-
hood 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 differ-
ent genes is systematically merged [19]. Third, the co-occurrence (this has also been coined phylogenetic
profile or COG pattern) of two genes (and their ortho-
logues) 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
generic 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 differ-
ences 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 informa-
tion 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 nat-
urally occurring mutations with known phenotype
association might, in some cases, substitute for experi-
ments 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://hbase.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 non-coding 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, allelic-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 of proteins.
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 α1-antichymotrypsin, a protease inhibitor of the serpin superfamily, that explain the loss of activity of plasma α1-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.
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 natural genomics. 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.

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


