Computational analysis of human disease-associated genes and their protein products
Kodangattil R Sreekumar, L Aravind and Eugene V Koonin*

The complete genome sequences for human, Drosophila melanogaster and Arabidopsis thaliana have been reported recently. With the availability of complete sequences for many bacteria and archaea, and five eukaryotes, comparative genomics and sequence analysis are enabling us to identify counterparts of many human disease genes in model organisms, which in turn should accelerate the pace of research and drug development to combat human diseases. Continuous improvement of specialized protein databases, together with sensitive computational tools, have enhanced the power and reliability of computational prediction of protein function.

Introduction
Inherited diseases are caused by mutation(s) in one or more genes. Diseases that are due to defect(s) in a single gene are called monogenic diseases; polygenic diseases are caused by defect(s) in more than one gene, with all those genes individually contributing to the development of the disease. In some diseases, such as Lafora’s (see below), the defect in proteins implicated in the disease can be directly related to the specific pathological state, whereas in others the link between the disease manifestation and the apparent defect observed at the level of the nucleotide or deduced amino-acid sequence may not be obvious. Irrespective of whether there is an obvious relationship between a defective gene and the pathological state, once a disease-associated gene (or a ‘disease gene’, in a common, if not precise, parlance) is identified, computational analysis of the gene is a critical step that can provide clues to the molecular basis of pathogenesis and invaluable insights for further experimental analysis.

Computational sequence analysis draws its strength from the conservation of critical features of a gene/protein in phylogenetically divergent sources. This branch of biology has greatly benefited from the finished and continuing genome-sequencing projects, as well as sequencing efforts undertaken on a smaller scale. Genome sequencing of three eukaryotic organisms — the fruitfly Drosophila melanogaster, the thale cress Arabidopsis thaliana and (in a draft form) Homo sapiens — has been completed in the past two years [1••–4••]. Table 1 lists some of the genes implicated in human disorders during the past year, and some of the previously identified human disease genes for which computational analysis in the same time frame has produced important new findings.

Procedures for undertaking rigorous sequence analysis, certain pitfalls in such analysis, methods to circumvent these problems, and the predictive power of computational analysis have been described previously [5–7]. Selected tools and databases that are currently available and are widely used in sequence and structure analysis are listed in Table 2. Here we focus on recent insights obtained by computational analysis of disease genes, and in particular how defects in one or more protein domains affect the cellular function of the encoded protein. The work that we discuss serves only to illustrate how computational analysis of disease genes can provide important clues about molecular basis of pathogenesis (we apologize to those researchers whose important contributions to this burgeoning field are not be cited owing to space limitations).

Conserved globular domains and their defects
Most proteins consist of one or more evolutionarily conserved domains, each with a distinct structure and function. An alteration of the amino-acid sequence of a domain is likely to hamper its proper functioning, especially if the alteration is in a highly conserved region and/or involves a non-conservative replacement of an amino-acid residue. Several tools are available for the automatic detection of protein domains to aid in function prediction (Table 2); however, the existing domain databases are still far from comprehensive, and the detection methods have limited sensitivity. It is therefore critical to analyze protein sequence in detail, on a case-by-case basis, by using several tools and methods and by assessing the relevance of the results obtained by computational approaches in the context of the available experimental data. Below, some recent discoveries of defects in proteins that result in human disorders are discussed with a view to relate a defective domain/motif to specific biological consequences.

Bcl10 and mucosa-associated lymphoid tissue lymphoma
A frameshift mutation in Bcl10, resulting in truncation distal to the caspase recruitment domain (CARD), is
Table 1

<table>
<thead>
<tr>
<th>Disease</th>
<th>Protein</th>
<th>Domain(s)* Other sequence features/motifs</th>
<th>Demonstrated or predicted protein function</th>
<th>Effect of mutation(s)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Soluble, globular proteins</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MALT lymphoma</td>
<td>Paracaspase</td>
<td>DEATH + 2(IG) + paracaspase</td>
<td>Predicted protein oligomerization, interaction with other proteins and protease activity potentially involved in apoptosis</td>
<td>Translocation produces a chimeric protein that might be an inhibitor of apoptosis.</td>
<td>[13]</td>
</tr>
<tr>
<td>MALT lymphoma</td>
<td>Bcl10</td>
<td>CARD Serine/Threonine-rich region</td>
<td>The Ser/Thr-rich domain may be involved in regulatory phosphorylation. The CARD domain mediates specific protein-protein interactions in the apoptotic system</td>
<td>Frameshift and truncation beyond the CARD domain might render the protein unstable or disrupt Ser/Thr phosphorylation</td>
<td>[8,9]</td>
</tr>
<tr>
<td>Fukuyama type CMD</td>
<td>Fukutin</td>
<td>Predicted phosphorysugar/choline transferase</td>
<td>Similarity to bacterial proteins suggests a role in modifying cell-surface glycoproteins or glycolipids</td>
<td>Transposon insertion disrupts the predicted enzymatic domain</td>
<td>[27,29]</td>
</tr>
<tr>
<td>Friedreich's ataxia</td>
<td>Frataxin</td>
<td>Frataxin Mitochondrial import peptide</td>
<td>Nuclear-encoded mitochondrial protein with a key role in the regulation of energy conversion</td>
<td>Several point mutations affect conserved amino-acid residues</td>
<td>[30,33]</td>
</tr>
<tr>
<td>Parkinson's disease</td>
<td>Parkin</td>
<td>Ubiquitin+ PARKIN_finger + RING</td>
<td>E3 ubiquitin-protein ligase</td>
<td>Mutations in ubiquitin domain affect binding to target proteins; mutations in RING prevent recruitment of E2</td>
<td>[55,56]</td>
</tr>
<tr>
<td>Giant axonal neuropathy</td>
<td>Gigaxonin</td>
<td>POZ+IVR+ kelch repeats</td>
<td>POZ-kelch domain combination suggests a role in cytoskeleton structure-assembly</td>
<td>Mutations that disrupt POZ or Kelch domains might disrupt protein-protein interactions causing in cytoskeleton defects</td>
<td>[57]</td>
</tr>
<tr>
<td>Retinitis pigmentosa</td>
<td>MERTK</td>
<td>Ig-like+ Fn3-like+ Tyr-kinase</td>
<td>Fn3 and Ig-like modules indicate ligand-binding ability; Tyr-Kinase domain implies protein phosphorylation; MERTK may be a receptor for a specific extracellular ligand</td>
<td>Mutations affecting Tyr-kinase domain disrupt the phosphorylation cascade activated by it.</td>
<td>[58]</td>
</tr>
<tr>
<td>Laterality defects</td>
<td>CFC1</td>
<td>EG F+ CFC Signal peptide and membrane-associating region</td>
<td>Extracellular signal molecule</td>
<td>Mutations in the conserved residues in EG F domain could disrupt inter-actions with its receptor</td>
<td>[59]</td>
</tr>
<tr>
<td>Macular corneal dystrophy</td>
<td>CHST6</td>
<td>Sulfo-transferase Signal peptide</td>
<td>Transfers sulfate groups from PAPS to various substrates</td>
<td>R50C mutation in the conserved region encompassing the sulfate donor binding site may prevent PAPS binding</td>
<td>[60]</td>
</tr>
<tr>
<td>Dominant optic atrophy</td>
<td>OPA1</td>
<td>Dynamin-like GTPase Mitochondrial import peptide</td>
<td>Possible role in maintenance and inheritance of mitochondria</td>
<td>R290Q, G300E, deletion of invariant I432 and a nonsense mutation in the dynamin GTPase domain</td>
<td>[45,46]</td>
</tr>
<tr>
<td>X-linked congenital SNB</td>
<td>nyctalopin</td>
<td>LRR-NT+15(LRR)+ LRR-CT Signal peptide, G P anchor</td>
<td>Predicted LRR-containing glyco-protein of extracellular matrix mediating protein-protein interactions</td>
<td>Missense mutations cause truncation, insertion and deletions affecting LRRs and/or loss of GP1-anchor</td>
<td>[61,62]</td>
</tr>
<tr>
<td>X-linked mental retardation</td>
<td>ARHGEF6</td>
<td>C H+SH3+ RhoG EF+ PH -</td>
<td>Signal transduction via the RhoGTPase cycle (RhoG EF domain); predicted to interact with actin filaments proline-rich peptides in proteins and inositol phosphate.</td>
<td>Intronic mutation resulting in exon 2 skipping, affects CH domain</td>
<td>[63]</td>
</tr>
<tr>
<td>Lafora's disease</td>
<td>EPM2A</td>
<td>CBD-4 +tyrosine phosphatase (PTP)</td>
<td>Binds polysaccharides via CBD-4; phosphatase domain may signal the catabolism of Lafora bodies.</td>
<td>W32G mutation in CBD, T194I in the PTP domain</td>
<td>[23]</td>
</tr>
</tbody>
</table>
## Table 1 (continued)

### Protein products of some genes mutated in human diseases.

<table>
<thead>
<tr>
<th>Disease</th>
<th>Protein</th>
<th>Domain(s)*</th>
<th>Other sequence features/motifs</th>
<th>Demonstrated or predicted protein function</th>
<th>Effect of mutation(s)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Usher syndrome type 1c</td>
<td>Harmonin</td>
<td>2(PDZ)+bZIP+PDZ</td>
<td>Proline/Serine/Threonine-rich region; leucine zipper</td>
<td>Predicted to dimerize via leucine zipper and mediate protein-protein interactions via PDZ domain. The P/S/T-rich region might serve as binding site for SH3 and WW domain-containing proteins</td>
<td>Mutations resulting in loss of most oral of the globular domains.</td>
<td>[64]</td>
</tr>
<tr>
<td>Familial segmental glomerulosclerosis</td>
<td>ACTN4</td>
<td>2(CH)+4(SPEC) + 2(EF)</td>
<td>-</td>
<td>Predicted actin-binding and cross-linking protein (CH domains) and Ca-binding (EF hands) protein. Probable component of the cyto-skeletal network (SPEC domain)</td>
<td>K228E, T232I, and S235P mutations in the 2nd CH domain cause increased affinity for actin</td>
<td>[65]</td>
</tr>
<tr>
<td>May-Hegglin anomaly and Sebastian syndromes</td>
<td>MYH9</td>
<td>Myosin + IQ</td>
<td>-</td>
<td>Nonmuscle myosin heavy chain 9; predicted to bind calmodulin (IQ motif)</td>
<td>N93K predicted to destabilize the 2nd helix of myosin catalytic domain, R70C predicted to alter helical stability and ATPase activity.</td>
<td>[66]</td>
</tr>
<tr>
<td>Mullibrey nanism</td>
<td>MUL</td>
<td>RING + B-box + BBC + MATH</td>
<td>-</td>
<td>Predicted nuclear protein, may participate in ubiquitin-mediated protein degradation, as a E3 ubiquitin ligase (RING) and possibly in apoptosis (MATH)</td>
<td>Deletions (splice site and loding) leading to truncation, with the loss of MATH domain in two cases</td>
<td>[41*]</td>
</tr>
<tr>
<td>McKusick-Kaufman syndrome</td>
<td>MKKS</td>
<td>Group II chaperonin</td>
<td>-</td>
<td>Facilitates correct protein folding in conjunction with ATP hydrolysis</td>
<td>One mutation predicted to affect intermolecular interaction based on structural modeling</td>
<td>[34**]</td>
</tr>
<tr>
<td>combined pituitary hormone deficiency</td>
<td>LHX3</td>
<td>2(LIM) + HOX</td>
<td>-</td>
<td>Predicted transcriptional regulator via DNA-binding homeodomain–protein interaction via LIM domain</td>
<td>A conserved Y replaced by C in LIM domain; frameshift leading to loss of homeodomain</td>
<td>[67]</td>
</tr>
<tr>
<td>Netherton syndrome</td>
<td>SPINK5</td>
<td>15 (KAZAL) Signal peptide</td>
<td>-</td>
<td>Potential extracellular adhesion molecule</td>
<td>Insertions and deletions leading to frameshift and protein truncation</td>
<td>[68]</td>
</tr>
<tr>
<td>familial cylindromatosis</td>
<td>CYLD</td>
<td>3(CAP-G-LY)+B-BOX+UCH-2</td>
<td>Proline-rich region</td>
<td>Predicted to coordinate attachment of cellular organelles to microtubules via CAP-G-LY domain and down-regulate protein degradation via the ubiquitin hydrolase via UCH-2 domain</td>
<td>Nonsense and splice site mutations leading to truncation of the C-terminal 2/3 of the protein</td>
<td>[69]</td>
</tr>
<tr>
<td>Type 2 diabetes</td>
<td>MAPK8IP1</td>
<td>SH3 + PTB</td>
<td>N-terminal low-complexity region, probably non-globular</td>
<td>Regulation of JNK signaling pathway by mediating protein-protein interactions via the SH3 and PTB domains</td>
<td>S59N mutation outside the two globular domains, not fully penetrant</td>
<td>[70]</td>
</tr>
<tr>
<td>Craniosynostosis and enlarged parietal foramina</td>
<td>MSX2</td>
<td>Homeobox</td>
<td>DNA-binding protein, transcription factor</td>
<td>DNA-binding protein, transcription factor</td>
<td>P148H enhances DNA-binding affinity of homeobox; R172H and deletion of R159-160 cause low DNA-binding affinity of homeobox</td>
<td>[15,16]</td>
</tr>
<tr>
<td>Tricho-rapholangeal syndrome type 1</td>
<td>TRPS1</td>
<td>8 (C2H2) + GATA + 2(C2H2)</td>
<td>-</td>
<td>Multiple DNA-binding domains predict a transcription factor</td>
<td>3 nonsense and 3 frameshift mutations cause loss of 2-4 Zn-fingers</td>
<td>[71]</td>
</tr>
<tr>
<td>Membrane proteins</td>
<td>EIF2AK3</td>
<td>β-propeller + TMS TM + S/T protein kinase</td>
<td>Signal peptide</td>
<td>ER membrane-associated translation initiation factor elf-2α kinase</td>
<td>Mutations in kinase domain or truncations N-terminal to the kinase domain. β-propeller domain in the N-terminus was previously undetected</td>
<td>[72]</td>
</tr>
</tbody>
</table>
associated with mucosa-associated lymphoid tissue (MALT) lymphomas that have the translocation t(1;14)(p22;q32) [8,9]. The CARD domain was first identified in a comparison of the sequences of several apoptotic proteins [10]; this CARD domain comprises six antiparallel α helices and mediates homotypic interactions between proteins involved in apoptosis [11]. Besides caspase recruitment, which is a critical step in the chain of events leading to apoptosis, protein–protein interactions mediated by CARD contribute to activation of the transcription

Table 1 (continued)

<table>
<thead>
<tr>
<th>Disease</th>
<th>Protein</th>
<th>Domain(s)*</th>
<th>Other sequence features/motifs</th>
<th>Demonstrated or predicted protein function</th>
<th>Effect of mutation(s)</th>
<th>Reference</th>
</tr>
</thead>
</table>

Disease products of some genes mutated in human diseases.

<table>
<thead>
<tr>
<th>Disease</th>
<th>Protein</th>
<th>Domain(s)*</th>
<th>Other sequence features/motifs</th>
<th>Demonstrated or predicted protein function</th>
<th>Effect of mutation(s)</th>
<th>Reference</th>
</tr>
</thead>
</table>

*Domain name abbreviations are as in the SMART database. The number of domains of a particular type is indicated (for example, 2IG denotes two consecutive immunoglobulin domains). Consecutive domains are connected with '+'.
factor NF-κB. The mutated Bcl10 protein seen in MALT lymphomas has lost its pro-apoptotic functions but retains the ability to activate NF-κB [12].

In a different MALT lymphoma translocation, t(11;18(q21;q21), a fusion of the inhibitor of apoptosis (IAP)-2 gene to the MLT1/MALT1 locus generates a chimeric protein comprising BIR repeats of IAP-2 and the caspase-like predicted protease domain of the protein designated paracaspase [13**]. The paracaspase family of caspase homologs has been identified recently, along with the metacaspase family, in a detailed computational analysis of the caspase-like protease superfamily [13**].

Paracaspases contain a predicted caspase-like proteolytic domain, a Death domain capable of mediating specific protein–protein interactions and, in certain cases, including the human representative immunoglobulin domains. In the MALT lymphoma translocation t(11;18(q21;q21), the prodomain of human paracaspase is replaced with BIR repeats that might function as inhibitors of apoptosis. The BIR–paracaspase fusion has been found to activate NF-κB in a manner dependent on the predicted catalytic cysteine of the paracaspase, thus supporting the computational prediction [13**]. Notably, the prodomain of human paracaspase interacts with Bcl10, which suggests that the two translocations associated with MALT lymphomas affect the same apoptotic pathway.

Alzheimer disease — presenilins
Familial Alzheimer disease is associated with mutations in genes that encode the paralogous integral membrane proteins presenilin 1 and presenilin 2. Presenilins are required to cleave several other integral membrane proteins, including the β-amloid precursor proteins Notch and Ire1 that are central to the pathogenesis of Alzheimer disease. It has been shown that two aspartate residues of presenilin 1 — one located in the intracellular loop between helices 6 and 7 and the other one located in helix 7 — are required for these proteolytic events, leading to the hypothesis that presenilins themselves belong to a distinct class of membrane proteases called γ-secretases [14**].

Sequence database searches using all available methods, including different types of profile analysis, have failed to detect similarity between presenilins and any known proteases. However, a pattern search carried out by Steiner et al. [14**] revealed that the signature (G[A]xGDb (where x is any residue, h is a bulky hydrophobic residue, and alternative residues are shown in brackets) is shared by presenilins and bacterial type-4 prepilin peptidases (TFPP), and includes one of the functionally critical aspartates of both protein families.

Presenilins and TFPPs have similar membrane topology, with eight transmembrane helices present in each family, but show no appreciable sequence similarity beyond the above signature. Therefore, although Steiner et al.’s [14**] observation reinforces the hypothesis that presenilins are the catalytically active γ-secretases, rather than cofactors of a still unidentified protease, it remains unclear whether they are homologs of TFPPs or whether the similarity in the (predicted) active sites is due to convergence. Given the relatively relaxed functional constraints that are typical of membrane proteins, with selection preserving mainly the membrane topology rather than sequence, common origin cannot be ruled out despite the absence of sequence conservation.

MSX1 and MSX2
Craniosynostosis and enlarged parietal foramina are caused by mutations in the homeodomain — a highly conserved DNA-binding domain containing three helical regions — of the MSX2 protein [15,16]. A mutation (P148H) that leads to craniosynostosis enhances the DNA-binding affinity of this protein, whereas another mutation (R172H) that results in parietal foramina has the opposite effect of lowering the protein’s affinity for DNA. Similarly, a mutation in the MSX1 gene (R31P) that affects the second helix of the homeodomain [17], and a nonsense mutation that leads to a truncation at position 104 and a peptide lacking the entire homeodomain [18] are implicated in tooth agenesis.

Other examples of diseases caused by mutations in homeodomains include microphthalmia [19], amegakaryocytic thrombocytopenia and radio-ulnar synostosis [20]. The paired (PAX) domain, another DNA-binding module involved in transcription regulation, is defective in certain individuals diagnosed with oligodontia [21]. An enhanced or diminished DNA-binding capacity of a transcription factor is likely to result in altered level of protein(s) encoded by genes under its control, which in turn might lead to pathogenesis.

Laforin
The presence of polyglucosan inclusion bodies in the brain is characteristic of progressive myoclonus epilepsy or Lafora’s disease. The EPM2A product, laforin, which is defective in patients suffering from this disease, was initially designated as a protein tyrosine phosphatase (PTP) because of the presence of a consensus sequence characteristic of the catalytic site of PTPs [22]. Minassian et al. [23**] have carried out a detailed computational analysis resulting in significant insights about this protein and its probable link to the accumulation of polyglucosan inclusions. Analysis of protein domains in laforin using the Pfam database [24] revealed the presence of the carbohydrate-binding domain (CBD)-4 at the amino (N) terminus of the protein, in addition to the dual-specificity phosphatase domain, which belongs to a distinct subfamily of PTPs, at the carboxyl (C) terminus. Furthermore, two sequence motifs characteristic of the glucohydrolase family as documented in PROSITE [25] were detected.

On the basis of these findings and the mutations detected in the EPM2A gene, Minassian et al. [23**] proposed that normal laforin prevents the accumulation of polyglucosan...
Table 2
Selected tools and databases useful in sequence analysis, with an emphasis on those dedicated to disease genes*.

<table>
<thead>
<tr>
<th>Tools/database</th>
<th>Comments</th>
<th>URL</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Tools and databases for functional annotation</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>COG</td>
<td>Phylogenetic classification of proteins from complete genomes that groups orthologous proteins; useful for functional annotation of newly sequenced genomes. COGNITOR can be used to place a protein into one of the COGs.</td>
<td><a href="http://www.ncbi.nlm.nih.gov/COG/">http://www.ncbi.nlm.nih.gov/COG/</a></td>
<td>[52+]</td>
</tr>
<tr>
<td>Pfam</td>
<td>Collection of protein domains and families consisting of multiple alignments and hidden Markov models generated in a semi-automatic manner.</td>
<td><a href="http://www.sanger.ac.uk/Pfam/">http://www.sanger.ac.uk/Pfam/</a></td>
<td>[24]</td>
</tr>
<tr>
<td>BLOCKS</td>
<td>Database of highly conserved regions of proteins derived automatically and represented in the form of ungapped multiple alignments. Also tools to detect and verify protein sequence conservation.</td>
<td><a href="http://www.blocks.fhcrc.org/">http://www.blocks.fhcrc.org/</a></td>
<td>[91]</td>
</tr>
<tr>
<td>PRINTS</td>
<td>Database of protein fingerprints (group of motifs distributed in the sequence characteristic of a family).</td>
<td><a href="http://bmbsgi11.leeds.ac.uk/bmb5dp/prints.html">http://bmbsgi11.leeds.ac.uk/bmb5dp/prints.html</a></td>
<td>[92]</td>
</tr>
<tr>
<td><strong>Tools for sequence similarity search</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BLAST</td>
<td>Programs for rapid sequence similarity searches for protein and DNA sequences. Also provides searches for nucleotide sequences translated in all six frames.</td>
<td><a href="http://www.ncbi.nlm.nih.gov/BLAST/">http://www.ncbi.nlm.nih.gov/BLAST/</a></td>
<td>[28]</td>
</tr>
<tr>
<td>HMMER</td>
<td>Programs for constructing multiple protein sequence alignments and performing database searches using the HMM approach.</td>
<td><a href="http://hmmer.wustl.edu">http://hmmer.wustl.edu</a></td>
<td>[93]</td>
</tr>
<tr>
<td>PSI-BLAST</td>
<td>An implementation of BLAST capable of detecting distantly related proteins by creating a PSSM from the significant matches detected in one round and using that PSSM as the query in the next iteration.</td>
<td><a href="http://www.ncbi.nlm.nih.gov/blast/psiblast.cgi">http://www.ncbi.nlm.nih.gov/blast/psiblast.cgi</a></td>
<td>[28]</td>
</tr>
<tr>
<td><strong>Tools for protein structure prediction and comparison</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PSI-PRED</td>
<td>Prediction of secondary structure using PSI-BLAST derived scoring matrices.</td>
<td><a href="http://insulin.brunel.ac.uk/psipred">http://insulin.brunel.ac.uk/psipred</a></td>
<td>[95]</td>
</tr>
<tr>
<td>FSSP/DALI</td>
<td>Automatic structural classification of protein domains with the DALI search facility for structure comparisons.</td>
<td><a href="http://www2.ebi.ac.uk/dali/fssp/">http://www2.ebi.ac.uk/dali/fssp/</a></td>
<td>[97]</td>
</tr>
<tr>
<td><strong>Databases for disease genes</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genes and Disease</td>
<td>Information on selected human diseases with links to genes, chromosomes, scientific literature and OMIM.</td>
<td><a href="http://www.ncbi.nlm.nih.gov/disease/">http://www.ncbi.nlm.nih.gov/disease/</a></td>
<td></td>
</tr>
</tbody>
</table>
inclusion bodies in neurons by binding polyglucosan through the CBD and cleaving the β-linkages through the putative glucohydrolase domain. A more detailed examination of the sequence and domain architecture of laforin shows, however, that the putative glucohydrolase domain largely overlaps with the confidently predicted dual-specificity phosphatase domain, which suggests that the presence of the glucohydrolase motifs in this protein is spurious (KR Sreekumar, L Aravind, EV Koonin, unpublished data).

The dual-specificity phosphatase activity of laforin has been confirmed experimentally [26]. The CBD domain of laforin probably targets the protein to polyglucosan inclusion bodies, whereas the phosphatase domain might activate a protein dephosphorylation pathway that triggers its destruction through a still unknown mechanism. This example illustrates both the utility of computational analysis of protein domains for function prediction and the need for caution in interpreting computational results.

**Fukutin**

Fukuyama type congenital dystrophy is caused by retroposon insertions in the gene encoding a protein called fukutin [27]. Extensive database searches using the PSI-BLAST program [28] detected moderate, but statistically significant similarity between fukutin and a family of bacterial and eukaryotic enzymes that catalyze phosphoryl-ligand transfer. Additional multiple alignment analysis showed that fukutin shares with these proteins the predicted catalytic residues [29].

The combination of these observations with the prediction of a signal peptide in fukutin, and the experimental data on its localization in the endoplasmic reticulum and secretory granules [27] has led to the prediction that fukutin modifies cell-surface molecules, most probably through the attachment of phosphoryl-sugar moieties [29].

**Frataxin**

Friedreich’s ataxia (FRDA) is an autosomal recessive disease characterized by progressive ataxia, hypertrophic cardiomyopathy and diabetes mellitus. Gibson et al. [30] have carried out computational analysis of the FRDA gene product, frataxin, and shown that it has homologues in bacteria of the γ subdivision of the Proteobacteria, but not in any other prokaryotes, and also that it possesses a non-globular N-terminal domain predicted to function as a mitochondrial targeting peptide [30]. They predicted a distinct fold, consisting of a β sheet flanked by two long α helices, on the basis of a multiple alignment of the frataxin homologs, and further proposed that frataxin is anuclear encoded, and, accordingly, that FRDA is a ‘mitochondrial disease’.

Both the mitochondrial localization and the αβ sandwich structure predictions, which were based on computational analysis of frataxin, have been confirmed experimentally [31,32], and more recent studies have shown that frataxin is a key regulator of mitochondrial energy conversion [33].

**McKusick-Kaufman syndrome**

Mutations in a gene on chromosome 20 have been shown to cause McKusick–Kaufman syndrome (MKKS) and Bardet–Biedl syndrome (see also review by Sheffield et al., this issue, pp 317–321) — developmental anomalies that involve hydrometrocolpos, postaxial polydactyly and congenital heart disease [34••,35]. Sequence database searches show that the predicted protein encoded by the MKKS gene belongs to a family of group II chaperonins that promote protein folding in an ATP-dependent manner.

The three-dimensional structural model of the MKKS protein predicts that the Y37C mutation lies in the highly conserved loop region between helix 1 and strand 2 which is involved in interaction among subunits [34••]. Another mutation, H84Y, is predicted to lie in a region responsible for ATP hydrolysis. This appears to be the first description of a disease caused by a defect in a molecular chaperone — a finding that complements the data on the role of defects in transcription regulators in several diseases, as discussed above.

**Vascular endothelial growth factor receptor-3**

Another example of a human disease for which sequence analysis and three-dimensional structure modeling has provided insights into pathogenesis is primary lymphoedema caused by mutations in vascular endothelial growth factor receptor-3 (VEGFR-3) [36••]. The VEGFR-3 protein consists of six classical immunoglobulin domains, four immunoglobulin C2 domains, a signal peptide, a single transmembrane segment and an intracellular tyrosine kinase catalytic domain.

Four mutations that co-segregate with lymphoedema phenotype map within the tyrosine kinase domain. Finegold and co-workers [36••] have built a three-dimensional model of VEGFR-3, based on the crystal structure of VEGFR-2, from which functional implications of mutations observed in VEGFR-3 can be inferred. For example, the model shows that the G857R mutation (the last glycine of the conserved GXGXXG kinase motif) lies in a critical turn between β strands 1 and 2 and is expected to affect ATP-binding and catalysis severely.

**DTD, Pendrin and congenital chloride diarrhoea**

Diastrophic dysplasia/achondrogenesis type IB (DTD), Pendred’s syndrome and congenital chloride diarrhoea are...
caused by malfunctions of anion transporters of the sulfate transporter family [37,38]. Unexpectedly, it has been shown that the C-terminal cytoplasmic portion of these proteins shares a conserved domain named STAS (after sulfate transporter anti-sigma) with bacterial anti-sigma-factor antagonists, and this connection provides clues for the regulation of anion transporters [39]. (Anti-sigma factors prevent formation of active RNA polymerase holoenzyme by trapping the sigma factor. Anti-sigma antagonists bind the anti-sigma factors and prevent the formation of sigma-anti-sigma complex.) The STAS domain of the antisigma-factor antagonist binds GTP and has a weak GTPase activity [40], leading to the prediction that the cytoplasmic domain of the anion transporters might regulate transport through NTP binding [39].

Mulibrey nanism
Muscle–liver–brain–eye (Mulibrey) nanism is an autosomal recessive disorder that affects several tissues of mesodermal origin, which suggests that a highly pleiotropic gene is involved in this disease. The complex multidomain architecture of the recently cloned MUL gene product is compatible with this notion [41]. The MUL protein comprises a RING-finger domain, a B-box domain, a coiled-coil domain and a MATH domain. Avela et al. [41] suggest that MUL is likely to be involved in various protein–protein interactions that might be important in development regulation; however, the combination of domains present in this protein calls for more specific functional predictions.

There is growing evidence that the primary function of the RING domain is as the E3 component of the ubiquitin ligase cascade that facilitates the transfer of ubiquitin from the E2 protein to target proteins [42]. MATH is a versatile adaptor domain that mediates specific protein–protein interactions in the programmed cell death system, chromatin remodeling and other functional contexts [43]. Thus, MUL can be predicted to function as an E3 ubiquitin ligase, with additional regulatory interactions, possibly related to apoptosis, mediated by the MATH domain.

Mutations outside functional protein domains
Mutations outside globular domains and transmembrane segments of proteins, or outside the protein-coding part of a gene altogether often affect gene expression. Such mutations may be located in untranslated regions, promoters, introns and sequences encoding signal peptides. The effect of these mutations may manifest itself as a defective globular or transmembrane domain, however, owing to exon skipping, aberrant splicing and abnormal cellular localization.

Untranslated and non-coding regions
Sequence analysis can provide hints as to the molecular basis of defects caused by mutations in untranslated regions. For example, in individuals with severe protein C deficiency, who suffer from massive disseminated intravascular coagulation or neonatal purpura fulminans, promoter mutations in the protein C (PROC) gene have been identified. These mutations alter the consensus sequence for the transcription factor HNF-3-binding, resulting in reduced PROC promoter activity, and accordingly a reduced level of protein C [44].

A splice-site mutation in intron 9 of the OPA1 gene, which has been detected in individuals with autosomal dominant optic atrophy, leads to either exon-10 skipping or a frameshift with a premature stop codon [45]. The OPA1 product is a mitochondrial protein comprising an N-terminal mitochondrial localization signal and a dynamin GTPase domain. The defective protein produced by the mutant gene lacks a part of the GTPase domain [45,46].

Signal peptides
Duplications in exon 1 of the TNFRSF11A (RANK) gene that segregates with familial expansile osteolysis have been identified recently [47]. These mutations are short in-frame insertions in the hydrophobic core region of the RANK signal peptide, and affect proper cleavage of the signal peptide.

The RANK protein consists of four tumor-necrosis factor repeats and a transmembrane segment, besides the signal peptide that directs the protein to the secretory pathway. Failure to cleave the signal peptide might result in higher intracellular accumulation of defective RANK in compartments of the secretion pathway that could lead to a higher incidence of receptor self-association and increased RANK signal transduction [47]. An NF-κB responsive reporter assay of transfected cells has indeed suggested that mutations occurring in familial expansile osteolysis result in increased constitutive RANK signaling [47].

Conclusions and future directions
The examples of disease genes discussed here and in previous publications [6,48–50] illustrate the importance of sequence and structure analysis in predicting the molecular basis of pathogenesis, in guiding further work and in understanding the biology of the respective functional systems. Proteins for which predictions based on sequence analysis have been experimentally verified include frataxin and the MALT-associated paracaspase. Using the complete genome sequences, efforts are being made to assemble databases of orthologous genes from diverse organisms, and these databases are already proving to be valuable in functional annotation of genes [51,52].

The presence of orthologs for many of the human disease genes in model organisms such as mouse, fruitfly, nematode and yeast enable experimental verification of the predictions. In particular, apparent counterparts for 178 of 287 analyzed human disease genes have been identified in fruitfly by large-scale comparative genomics [53,54]. At present, some of the human disease genes do not seem to have an ortholog in other organisms, but the anticipated availability in the next several years of the complete genome sequences of primates and other mammals should drastically reduce the number of ‘orphan’ disease genes.
The predictive power of sequence and structure analysis, combined with the wealth of genome sequence data, should not only enhance our understanding of the biology that underlies the effect of mutations in disease genes, but also enable the identification of many new drug targets and accelerate the process of drug design and the development of therapeutic strategies.

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


The second (nearly) complete genome of a multicellular eukaryote to be sequenced, after the yeast *Saccharomyces cerevisiae*. The completion of the fly genome is particularly important: first, because the enormous wealth of genetic data available for this organism can be systematically correlated with gene and protein sequences; and second, because an in-depth comparison of two animal genomes has become possible for the first time.


Sequencing of the first complete plant genome is fundamentally important as it sets the stage for a comparative study of genomes representing the three major lineages of the eukaryotic crown group — animals, fungi and plants.


This paper reports the results of the public effort on human genome sequencing. A preliminary analysis of the predicted human proteome performed with a variety of computational techniques, with particular emphasis on the ‘Interpro’ collection of protein domains, is presented. The major conclusions are that human proteins consistently have more complex domain architectures compared to their homologs from other eukaryotes and domain rearrangements have a prominent role in eukaryotic evolution. Over 1300 apparentorthologs common to human, fly, worm and yeast have been identified.


The report of the human genome sequence from the private company Celera Genomics. In the preliminary analysis of the predicted human proteome, molecular functions for ~65% of the proteins have been assigned by automatic means utilizing the protein family databases such as Pfam and SMART. The authors have also identified a partial set of human–fly and human–wormorthologs. A preliminary survey of the differences between the human genome and other sequenced eukaryotic genomes is also presented.


This example of combining computational analysis and experimental verification reveals the molecular basis of a class of MALT lymphoma translocation. Computational analysis using sensitive tools for sequence-profile analysis was critical for the identification of two new families of caspase-related proteases, which produce a new perspective on the evolution of this important class of enzymes.


Reports the site-directed mutagenesis of presenilin 1 and detection of the similarity between its putative catalytic site and that of bacterial type-4 prepilin peptidases. A functional and possibly evolutionary connection, revealed by computer analysis of protein sequences and structures, is plausible despite the absence of significant sequence similarity.


This paper underscores both the potential of computational analysis of protein domains and the need to be cautious in interpreting such results. Computational analysis of laforin reveals the presence of a carbohydrate-binding domain and a dual specificity phosphatase domain; however, the apparent glycohydroxylase motifs detected in this protein are likely to be spurious.


Mutations in MYH9 result in the May-Hegglin anomaly, and

In Mutations in MYH9 result in the May-Hegglin anomaly, and

In Mutations in MYH9 result in the May-Hegglin anomaly, and

In Mutations in MYH9 result in the May-Hegglin anomaly, and

In Mutations in MYH9 result in the May-Hegglin anomaly, and

In Mutations in MYH9 result in the May-Hegglin anomaly, and

In Mutations in MYH9 result in the May-Hegglin anomaly, and

In Mutations in MYH9 result in the May-Hegglin anomaly, and

In Mutations in MYH9 result in the May-Hegglin anomaly, and

In Mutations in MYH9 result in the May-Hegglin anomaly, and

In Mutations in MYH9 result in the May-Hegglin anomaly, and

In Mutations in MYH9 result in the May-Hegglin anomaly, and

In Mutations in MYH9 result in the May-Hegglin anomaly, and

In Mutations in MYH9 result in the May-Hegglin anomaly, and

In Mutations in MYH9 result in the May-Hegglin anomaly, and

In Mutations in MYH9 result in the May-Hegglin anomaly, and

In Mutations in MYH9 result in the May-Hegglin anomaly, and

In Mutations in MYH9 result in the May-Hegglin anomaly, and

In Mutations in MYH9 result in the May-Hegglin anomaly, and

In Mutations in MYH9 result in the May-Hegglin anomaly, and

In Mutations in MYH9 result in the May-Hegglin anomaly, and

In Mutations in MYH9 result in the May-Hegglin anomaly, and

In Mutations in MYH9 result in the May-Hegglin anomaly, and

In Mutations in MYH9 result in the May-Hegglin anomaly, and

In Mutations in MYH9 result in the May-Hegglin anomaly, and

In Mutations in MYH9 result in the May-Hegglin anomaly, and

In Mutations in MYH9 result in the May-Hegglin anomaly, and

In Mutations in MYH9 result in the May-Hegglin anomaly, and

In Mutations in MYH9 result in the May-Hegglin anomaly, and