Application of in silico positional cloning and bioinformatic mutation analysis to the study of eye diseases

Agnieszka M. Lichanska and David A. C. Simpson

Abstract

A vast amount of DNA and protein sequence is now available and a plethora of programs have been developed to analyse the data. The bewildering variety of analyses that can be performed via the World-Wide Web can deter researchers from applying bioinformatics to augment their traditional genetic research. Focusing on the inherited eye diseases, this paper provides a guide to the appropriate software required for identification of candidate genes through to the detection and analysis of mutations.

INTRODUCTION

The aim of this paper is to show how bioinformatics can be used to advance research into inherited eye diseases by making predictions that direct conventional approaches. We describe how to use bioinformatic tools to identify candidate genes, to discover and analyse mutations in patients and how to exploit this information to develop possible treatment strategies (Figures 1 and 2).

Bioinformatic approaches are being increasingly employed to facilitate cancer research (through functional genomics and proteomics), but it appears that they are under-utilised in other fields, ophthalmology being one such area. It is particularly amenable to such an approach, especially in the area of candidate gene identification and characterisation of mutations detected in ocular diseases. There is a range of heterogeneous hereditary diseases that affect all ocular tissues, in particular retina, lens, cornea and vitreous. Their phenotype can range from mild to severe visual impairment, from congenital to midlife blindness. The most difficult task facing researchers in ophthalmology is to identify the genes responsible for each disease and then to correlate the genotype to the clinical presentation. Mutations in the same gene can cause very different diseases. For example, mutations within single genes such as the peripherin/RDS or cone-rod homebox genes can cause at least three outer retinal disorders.\(^1\)\(^2\)\(^3\)

Moreover, a disease can have a number of very similar clinical manifestations but be genetically heterogeneous (eg central areolar choroidal dystrophy, CACD, or retinitis pigmentosa, RP).\(^1\)\(^4\)\(^5\)

Additionally, some diseases present a phenotypical spectrum, with strong genetic predisposition but no genetic loci as yet identified (eg macular degenerations, especially age-related macular degeneration, AMD).

Correlation between the genotype and phenotype is particularly complex in eye diseases and more than half of the genes responsible for the hereditary eye diseases where the locus has been mapped have not yet been cloned (RetNet,\(^6\) OMIM\(^7\)). This is hardly surprising considering that the disease genes are often linked to
regions over 2Mb, which can contain over 120 genes, most of which are not yet characterised. However, the identification of the gene involved in a specific ocular disorder and the changes induced by mutations within it are essential as this knowledge will form the basis for future therapeutic strategies.

**Figure 1**: A schematic outline depicting the contribution of bioinformatics to the investigation of genetic eye disease. Improvements in genetic maps enable linkage of disease genes to smaller genomic regions and the existence of comprehensive nucleotide sequence data enables prediction of the genes within the region of interest. The proliferation of EST sequences (arrows) facilitates prediction of gene structure and whether specific transcripts are expressed within the eye. Analysis of mutation databases can indicate changes within candidate genes and preliminary predictions of their effect on protein structure/function may indicate the likelihood of their involvement in a disease phenotype.

**Figure 2**: The role of bioinformatics in the identification and analysis of mutations. Many steps in the progression from initial linkage analysis to determination of a mutation responsible for a specific disease can now be performed or greatly aided by a bioinformatic approach. These steps are enclosed in the shaded box, while the processes still requiring ‘wet’ laboratory analyses are indicated outside the box.
IDENTIFICATION OF A DISEASE GENE
Defining of a chromosomal region and determination of gene content within this region

Many of the inherited eye diseases have been mapped by linkage analysis to a chromosome but a big step is now needed to identify the gene that when mutated is responsible for the disease. With a rapid increase in the amount of sequence data available from the human genome project it will soon be possible to replace laboratory-based positional cloning with in silico identification of candidate genes (Figures 1 and 2). This will eliminate the time-consuming sequencing and contig creation stage.

With the Human Genome Project nearly completed it is now possible to identify most of the genes within a particular chromosomal region by searching physical maps of draft and finished sequences. These searches can be performed using Map Viewer[10] or Ensembl, Sanger Center, EBI[11]. In this way, genomic sequences within the region of interest can be identified. For some time gaps will exist in the human genome and therefore additional library screening and sequencing may be required to obtain full sequence across the region of interest.

Both tools are in principle similar, the main differences being in the presentation of the data and annotation system. Map Viewer primarily enables the identification of the genomic contigs within a region of interest, but also contains links to all genes identified within the genomic region, their structure and conserved domains within the protein sequence. Navigation through the genome can be difficult because of awkward navigation and zooming systems, and only limited information on the region is immediately available.

In contrast, Ensembl provides all the information available for the region within the viewing screen, allows easy zooming in and out of the region viewed as well as navigation across the chromosome. It also offers partial annotation of the sequence by incorporating existing genes and genes identified by computational tools for gene structure prediction (GeneWise or Genscan). Links are provided to the GenBank, SNP, EST, Unigene and SWISS-PROT databases, and to evidence supporting the structure of computationally predicted genes.

Evidence for the predictions is based on sequence similarity of the genomic fragment to sequences of ESTs, RIKEN clones, UniGene sequences and SWISS-PROT protein sequences. This raises the issue of whether the genomic fragments have been correctly assembled, as any changes in assembly will affect the predictions. Therefore, independent confirmation of the gene structure is essential.

In general, Ensembl has a more user-friendly interface and by collating all the information available on the sequences within the viewing region it allows easy identification of the relevant information.

Computer-based prediction of gene structure

It is important to bear in mind that a number of genes incorporated into the genome browsers were computationally predicted. This results in listing of hypothetical genes, which may in reality represent pseudogenes. Conversely, real genes may be missed due to limited ability to predict, for example, intronless genes, overlapping genes and multiple genes in a single sequence (especially on the same strand).[12] Tools commonly used for prediction of gene structure can be divided into three groups: splice site prediction, exon identification and full gene prediction (reviewed recently by Thanaraj[13]). Identification of an entire gene within a genomic fragment relies on the programs (eg HHMgene, GeneScan) to recognise translational signals (Kozak sequence and in-frame termination codon) in addition to identifying exons by recognising splice sites. It appears that the ability to identify initial and terminal

© HENRY STEWART PUBLICATIONS 1467-5463. BRIEFINGS IN BIOINFORMATICS. VOL. 3. NO 1. 59-72. MARCH 2002
**Use of multiple gene prediction tools gives a more reliable gene model**

Exons vary between and within programs, as most programs (except for GenLang and GeneScan) have greater difficulty in identifying terminal rather than initial with 20–43 per cent of terminal ones missed. In contrast, identification of internal exons only (Grail, MZEF) relies mostly on the ability of the program to predict the splice sites and although these are well conserved (GT at 5′ and AG at 3′ end), various other factors independent of the algorithms themselves can contribute to observed problems, including misassembly, incomplete or inaccurate sequence or presence of unusual sequences such as, for example, repetitive motifs within exons (A.M.L., unpublished observations). In summary, the accuracy of exon prediction varies widely from 40 to 90 per cent depending on the tool used. Detailed reviews on gene prediction software have been published elsewhere and are useful for in-depth analysis of the programs available.

An important issue in analysis of uncharacterized fragments of genomic sequence is the decrease in accurate predictions in novel sequences (especially large ones). This problem is considered to be caused by the training sets used in development of the programs as they rely on often artificially short fragments of high coding density. As a result, when programs are confronted with large genomic sequences, their predictive power is low and results unreliable, as illustrated recently by the comparison of the genes predicted by Celera and Ensembl. It is estimated that nearly 80 per cent of novel transcripts are predicted by only one of the groups in comparison with approximately 84 per cent agreement with the reference sequences in the database. The main reasons for such large discrepancies may be incomplete or inaccurate sequence assembly, algorithms used in predictions, incomplete transcriptome representation in the database and presence of a large number of RNA splice variants. It will take some time to fully resolve which predictions are correct. Owing to all the problems with gene structure predictions attempts have been made to create programs aimed at eliminating the false positives. An alternative approach (for small-scale analysis) is to use a more complex sequence analysis tool, NIX at the HGMP-UK site, which enables simultaneous evaluation of the predictions from multiple programs (Table 1) with different predictive abilities (Fex, Hexon, MZEF, Genemark, Grail, Genefinder, Fgene, GeneScan, Fgenes), giving the user opportunity to compare identified exon and gene structures. The conclusion of this discussion is that no prediction is 100 per cent accurate and therefore benchmark is essential to confirm the predictions.

**CONFIRMATION OF A GENE STRUCTURE**

The confirmation of a gene structure is, however, aided by the presence of a large number of IMAGE cDNA clones that can be used in the same way as cDNA clones obtained by regular screening of cDNA libraries. There are also other library resources such as RIKEN, and a number of commercial sources (eg Invitrogen, Stratagene), but the main advantage of the IMAGE set is its free availability to

---

**Table 1: Gene structure prediction tools contained in NIX analysis tool**

<table>
<thead>
<tr>
<th>Tool</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grail</td>
<td><a href="http://avalon.epm.ornl.gov/Grail/bin/EmptyGrailForm">http://avalon.epm.ornl.gov/Grail/bin/EmptyGrailForm</a></td>
</tr>
<tr>
<td>GeneScan</td>
<td><a href="http://bioweb.pasteur.fr/seqanal/interfaces/genscan.html">http://bioweb.pasteur.fr/seqanal/interfaces/genscan.html</a></td>
</tr>
<tr>
<td>Genemark</td>
<td><a href="http://genemark.biology.gatech.edu/GenMark/">http://genemark.biology.gatech.edu/GenMark/</a></td>
</tr>
<tr>
<td>Genefinder, MZEF</td>
<td><a href="http://sciclio.cshl.org/genefinder/">http://sciclio.cshl.org/genefinder/</a></td>
</tr>
<tr>
<td>Fgene, Fex, Hexon</td>
<td><a href="http://genomic.sanger.ac.uk/gfif.html">http://genomic.sanger.ac.uk/gfif.html</a></td>
</tr>
</tbody>
</table>
researchers world-wide. RIKEN, although having the advantage of full-length clones, requires licensing before the clones are released and at the moment only a full set of clones is available.

The sequences of these clones are being deposited in expressed sequence tag (EST) databases (Genbank, EMBL) and allow an immediate comparison with genomic sequence to determine intron/exon boundaries without the need for any library screening, cloning and sequencing. The most common problem with EST analysis is poor sequence quality, while occasionally clones may be chimaeric or contain genomic sequence. Moreover, alternative splicing and partial sequencing means that clones from the same gene may not be clustered together. If an EST is identified within the region of interest a number of clones should be fully sequenced and then compared with the genomic sequence to prevent potential misinterpretation of genomic structure. Such computerised analysis of sequences can encounter problems if the clones represent different splice forms of the same gene, which is especially likely if the clones are derived from different tissues. This problem can be easily avoided if the full EST sequences are firstly compared with each other before analysis against the genomic sequence. Once the genes in a region have been defined, their expression in the eye can be checked by analysing whether any of the ESTs representing a specific gene are derived from cDNA libraries prepared from ocular tissue. However, the under-representation in the EST database of sequences from ocular libraries can lead to false negatives in this analysis. The searches can be performed by searching the UniGene division of GenBank or even more conveniently by using a new tool, GetMaps, which interprets the expression information stored in the UniGene database for genes expressed in the retina. Despite the limited number of retinal ESTs, such computer-based analysis can still generate a number of useful candidate genes which are expressed in the eye for further analysis.

An important step is to confirm that the cDNA clone is full length and is expressed in the retina. This can be performed by Northern blot analysis, RACE (rapid amplification of cDNA ends) analysis or in situ hybridisation (to identify the site of expression).

To prioritise which genes are to be screened first for mutations, additional analysis should be performed to try to elucidate the potential functions of these genes. At this stage one will identify two types of genes in the region of interest, well-known genes (eg ACADVL in the CACD locus) and novel genes (eg HPC002). In both cases it is quite likely that their role in the development and biology of the human eye is unknown. Although for the well-known genes homologues might have been described in other animals, such information is not available for the novel genes. In order to determine similarity of these proteins to known ones BLAST and position-specific iterating BLAST (PSiBLAST) database searches can be used to identify functionally related proteins in other species that might have been well characterised. Although some of the conserved domains will be identified in this manner, the sensitivity of BLAST is limited. More information can be obtained by searching profile/pattern databases using Pfam, Prosite or SMART, as recently reviewed for matches which can point to a possible function of the novel protein. These tools can reveal a more complete picture of the domain structure of a novel protein, providing a better indication of possible function.

Presence of functionally similar proteins in other species can be exploited even further by searching the mouse genome informatics site (MGI) to determine if the role of the protein in question has been identified in any of 19 species listed (primates, cattle, dogs, cats, rats, mice, rabbits, wallabies, platypus, etc.). Such searches are important for the identification of the best animal models.
Finding an animal model can aid in the development of therapies for a given disease. MGI provides information on tissue expression patterns, known mutations in a gene, probes and polymerase chain reaction (PCR) primers used for amplification of different alleles. Further functional information on the genes in the region of interest can be gained from existing transgenic animals by searching transgenic databases (Table 2). Tbase, in particular, contains detailed information on mutant mice, which can be useful in assessing how similar the pathology observed is to that in a human disease. Animal models can be used in two ways: firstly to look at the pathology caused by absence or mutation of the gene and secondly to evaluate possible treatments (e.g. rhodopsin13). There are, however, some limitations to the use of animal models. For example, while mice are reasonable models for retinal disorders they cannot be used for investigating macular diseases since mice do not have a macula and any phenotype observed will therefore be different from that in humans.

In short, we will soon be able to obtain the sequence of the region involved in a disease, determine its gene content and which genes are expressed in the retina mostly by computational means (Figure 2). At the moment there are, however, some shortcomings of the bioinformatic approach, mainly due to the incompleteness of the human genome sequence, under-representation of retinal ESTs in the database and inappropriate or incomplete annotation of the sequences in the databases and limitations of the algorithms used for DNA analysis. One cannot fully rely on the computational analysis of the human genome and therefore the retinal expression of a gene has to be confirmed by ‘wet’ methods such as in situ hybridisation, reverse transcriptase PCR (RT-PCR) or Northern blot analysis.

There are, however, steps that are still based upon benchwork, namely linkage analysis and confirmation of mutations in the candidate gene (Figure 2). The region within a gene that contains a mutation can be narrowed down by using various methods, including single strand conformational polymorphism (SSCP), heteroduplex analysis (HDA), denaturing gradient gel electrophoresis (DGGE) or wave analysis (using a transgenic DNA fragment analyser). Restriction enzyme mapping of genomic fragments, previously achieved by restriction fragment length polymorphism (RFLP) analysis, has been given a new twist with the development of virtual genome scan (VGS),14 a tool that predicts spots representing restriction fragments on restriction landmark genome scanning (RLGS) gels. At the moment this tool is unable to predict all fragments observed as the human genome sequence is still incomplete.

Once the region associated with a

### Table 2: Transgenic databases

<table>
<thead>
<tr>
<th>Database</th>
<th>Address</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tbase</td>
<td><a href="http://tbase.jax.org/">http://tbase.jax.org/</a></td>
</tr>
<tr>
<td>Transgenic and knockout mouse</td>
<td><a href="http://immunologylink.com/transgen.htm">http://immunologylink.com/transgen.htm</a></td>
</tr>
<tr>
<td>Mouse Knockout and Mutation Database</td>
<td><a href="http://research.bmim.org/mkmd/">http://research.bmim.org/mkmd/</a></td>
</tr>
<tr>
<td>UCD Medpath Transgenic Mouse Searcher 2.0^</td>
<td><a href="http://lena.jax.org/resources/documents/mmr/">http://lena.jax.org/resources/documents/mmr/</a></td>
</tr>
<tr>
<td>Skarnes Laboratory Resource for gene trap insertions</td>
<td><a href="http://socrates.berkeley.edu/~skarnes/resource.html">http://socrates.berkeley.edu/~skarnes/resource.html</a></td>
</tr>
<tr>
<td>Nagy Laboratory: Cre Transgenic and Floxed Gene Databases</td>
<td><a href="http://www.mshri.on.ca/nagy/cre.htm">http://www.mshri.on.ca/nagy/cre.htm</a></td>
</tr>
<tr>
<td>Transgenic rodents for the study of mutation in animals</td>
<td><a href="http://eden.ceh.unic.ca/bigblue.htm">http://eden.ceh.unic.ca/bigblue.htm</a></td>
</tr>
</tbody>
</table>

^Needs permission to use.
Choice of a reference sequence for analysis is very important

Changes in DNA can be pathogenic mutations or polymorphisms

In silico mutation generation

disease has been identified the nucleotide sequence must be determined. The first analysis of sequence obtained from the patient will be to compare it with the reference sequence for that gene. However, as reference sequences are not error-free\(^{35}\) it would be more effective to align a number of sequences of a gene from unaffected people to generate an in-house consensus/reference sequence. A subsequent comparison of the patient’s sequence with a consensus can be made using the BLAST 2 sequences program.\(^{36}\)

If more than two samples are being analysed, multiple alignment tools such as ClustalW,\(^{36,37}\) GCGPileup\(^{38}\) or GAP\(^{39}\) could be used to compare the sequences.

NATURE OF MUTATIONS

The identification of changes in a DNA sequence (mutations) is just the beginning of the analysis of the disease. These changes can involve the whole chromosomes (eg gain or loss of chromosomal fragments or rearrangements) or substitutions, insertions or deletions of single or multiple base pairs. Mutations can be either non-pathogenic (polymorphisms), giving rise to various alleles and phenotypic variability, or deleterious, leading to a disease (pathogenic variation in sequence). Changes predicted to affect splice junctions or alter one amino acid to another, cannot necessarily be assumed to be pathological without further evidence (eg RT-PCR analysis for potential splice mutations or functional analysis of a mutated protein). In order to confirm that an observed change is a pathologial mutation rather than a rare polymorphism, allele frequency in a normal population should be determined by using ethnically matched normal controls.

Polymorphisms occur naturally at high frequencies in a population and can include single nucleotide changes (SNPs) or changes in larger sequences (eg variable number of tandem repeat (VNTR) polymorphisms) and have the potential to change the protein sequence and structure but without deleterious effects on the organism. However, an accumulation of apparently non-pathogenic mutations can lead to predisposition to certain diseases. In contrast, disease-causing mutations in single gene disorders usually alter a gene product, affecting the survival or health of an organism. These changes can result in protein truncation or alterations in secondary and tertiary structure or/and subcellular localisation. A change in protein structure can also affect protein–protein or protein–ligand interaction.

Intriguingly, mutations in different parts of the same protein can cause varying clinical presentations.\(^{40-42}\) It is important to determine why a mutation causes a disease. The precise knowledge of the way a mutation affects the protein structure and protein interactions can greatly improve our understanding of the disease process and aid the design of possible treatments. A detailed computational analysis of the mutations responsible for eye diseases is very important because of limited access to the tissue itself and inability to obtain biopsy material from patients. It is, however, quite clear that some of the steps can involve a large amount of benchwork.

MUTATION ANALYSIS

The functional significance of a mutation has to be assessed with a conventional (laboratory-based) approach but can be assisted and directed by computational methods. Conventional functional analysis of mutants is a time- and labour-intensive process and can be very expensive if a large number of mutants have to be generated and screened. Computational methods can shorten the time required by modelling a large number of mutant proteins in silico. Bioinformatic methods are not a substitute for laboratory-based work but can provide the first look at the probable changes in protein structure and enable selection of the most interesting mutants for wet analysis. The mutants will then be created, confirmed for the presence of
mutation and their phenotype characterised by suitable methods.

METHODS FOR ANALYSING PROTEIN STRUCTURE
Secondary structure, subcellular targeting and protein modification
Depending on the type of mutation the impact on the final protein product will be different. Nucleotide variations resulting in frame-shift or truncation of the protein will be easily identified by BLAST comparison of the conceptual translation product with a normal sequence. The most difficult mutational effects to assess are changes caused in secondary folding, subcellular localisation or the post-translational modification pattern. There are multiple tools available on the Internet with different algorithms to investigate these alterations.

Changes affecting secondary folding might not be detected by the current tools because mutations do not change the structure sufficiently for the current algorithms to detect. Analysis of rhodopsin mutants (A.M.L., unpublished results) has shown that unless the change causes a significant break in the structure of a helix the alteration to folding might go undetected. If one wants to determine if there are any changes in the secondary structure and targeting sequences a good place to start is Jpred.3,44 This is a consensus method for protein structure prediction which combines multiple prediction methods (PHD, DSC (home server is currently unavailable), PREDATOR, NNSSP, MULPRED, ZPRED, JNET, JNETsolacc, COILS, MULTICOIL and PHDhtm – Table 3), thereby reducing the total time spent on analysis. To align the query protein with any available homologues, which improves secondary structure prediction, Jpred uses PSIBLAST, SCANPS55 and CLUSTALW46 alignment methods. The output gives the results for each of the programs used as they are run in parallel.

The accuracy of the prediction is estimated to be approximately 70 per cent. The advantage of using Jpred is that it allows direct comparison of various outputs and the consensus prediction provides a more accurate structural analysis than individual programs. The tools used for identification of transmembrane domains (Table 3) are slightly more sensitive, as they are able to identify the changes in length of transmembrane and loop domains based on their structure.

In analysing rhodopsin mutations we found that the domain screen of the Prosite database using ProScan at Pôle Bio-Informatique Lyonnais (PBIL) and Pfm (EBI)57 was very helpful in identifying the functional protein domains (eg opsins binding, myristylation sites, protein kinase C (PKC) phosphorylation) that are affected by changes in protein sequence. These searches identify most of the sites of post-translational modifications (eg phosphorylation). Additionally, alterations of N- and O-glycosylation and also glycosylphosphatidylinositol (GPI) anchor formation can be analysed using other programs (Table 3). The reliability of predicted modification sites and changes due to mutations is good, with the accuracy for GPI anchors reaching 100 per cent.

Targeting sequences form an important part of proteins and are required for them to be appropriately distributed within the cell. Although a number of algorithms attempt to predict subcellular localisation (Table 3), they are based on the probability of the protein being in one compartment rather than another. Therefore the results show the localisation in a particular organelle as a percentage rather than absolute prediction. Particular problems can be observed with proteins that belong to the secretory pathway within the cell, as their localisation can be predicted in multiple compartments, and in such cases distribution of the mutant protein will have to be assessed in living cells.
### Table 3: Tools for analysis of protein structure

<table>
<thead>
<tr>
<th>Secondary structure prediction</th>
<th>PHD</th>
<th><a href="http://www.embl-heidelberg.de/predictprotein/">http://www.embl-heidelberg.de/predictprotein/</a></th>
<th>On-line</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NNSP – nearest neighbour SS prediction; SSPE – segment oriented SS prediction; SPPAL – nearest-neighbor with local alignments SS and accessibility prediction; PSS – search for Prosite patterns with statistics</td>
<td><a href="http://genomic.sanger.ac.uk/pss/pss.shtml">http://genomic.sanger.ac.uk/pss/pss.shtml</a></td>
<td>On-line</td>
</tr>
<tr>
<td></td>
<td>Predator</td>
<td><a href="http://www.embl-heidelberg.de/cgi/predator_serv.pl">http://www.embl-heidelberg.de/cgi/predator_serv.pl</a></td>
<td>On-line</td>
</tr>
<tr>
<td></td>
<td>DSC (Prof)</td>
<td><a href="http://www.aber.ac.uk/~phi/www/prof/">http://www.aber.ac.uk/~phi/www/prof/</a></td>
<td>Executables are available for UNIX and DOS platforms from ftp.ebi.ac.uk</td>
</tr>
<tr>
<td></td>
<td>Coils</td>
<td><a href="http://www.ch.embnet.org/software/COILS_form.html">http://www.ch.embnet.org/software/COILS_form.html</a></td>
<td>On-line and for download by ftp</td>
</tr>
<tr>
<td></td>
<td>ZPRED</td>
<td><a href="http://kestrat.ludwig.ucl.ac.uk/zpred.html">http://kestrat.ludwig.ucl.ac.uk/zpred.html</a></td>
<td>On-line, multiple alignment from AMPS program as input</td>
</tr>
<tr>
<td></td>
<td>MULTICOIL</td>
<td><a href="http://nighengale.lcs.mit.edu/cgi-bin/multicoil">http://nighengale.lcs.mit.edu/cgi-bin/multicoil</a></td>
<td>On-line, Downloadable version for SUN, Silicon Graphics and DEC ALPHA</td>
</tr>
<tr>
<td></td>
<td>GOR III, GOR IV, HNN</td>
<td><a href="http://purl.ibcp.fr/cgi-bin/nsa_automat.pl?page=">http://purl.ibcp.fr/cgi-bin/nsa_automat.pl?page=</a></td>
<td>On-line</td>
</tr>
<tr>
<td></td>
<td>SOMP, SOMP</td>
<td>NPSA/npsa_server.html</td>
<td></td>
</tr>
<tr>
<td></td>
<td>NN Predict</td>
<td><a href="http://www.cmopharm.ucsf.edu/~nmomi/nnpredict.html">http://www.cmopharm.ucsf.edu/~nmomi/nnpredict.html</a></td>
<td>On-line</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Prediction tools for protein modification</th>
<th>Glycosylation</th>
<th><a href="http://www.cbs.dtu.dk/services/NetOGlyc">http://www.cbs.dtu.dk/services/NetOGlyc</a></th>
<th>60-63</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Phosphorylation</td>
<td><a href="http://www.cbs.dtu.dk/services/NedPhos">http://www.cbs.dtu.dk/services/NedPhos</a></td>
<td>64-65</td>
</tr>
<tr>
<td></td>
<td>GPI anchors</td>
<td><a href="http://rmdndl.imp.univie.ac.at/gpi/gpi_server.html">http://rmdndl.imp.univie.ac.at/gpi/gpi_server.html</a></td>
<td>65, 66</td>
</tr>
<tr>
<td></td>
<td>Mitochondrial targeting</td>
<td><a href="http://www.cbs.dtu.dk/services/SignalP">http://www.cbs.dtu.dk/services/SignalP</a></td>
<td>68-69</td>
</tr>
<tr>
<td></td>
<td>GenoMine</td>
<td><a href="http://www.bionformatics.uc.edu/genomine">http://www.bionformatics.uc.edu/genomine</a></td>
<td>Linux, Sun Solaris, Silicon Graphics</td>
</tr>
<tr>
<td></td>
<td>WHAT IF</td>
<td><a href="http://www.psimx.co.uk/products/naomi/">http://www.psimx.co.uk/products/naomi/</a></td>
<td>Silicon Graphics</td>
</tr>
<tr>
<td></td>
<td>Sybyl</td>
<td><a href="http://www.cmbi.kun.nl/1I00/WIVWWW/">http://www.cmbi.kun.nl/1I00/WIVWWW/</a></td>
<td>DEC-alpha, Sun, IBM PC running Linux or DOS</td>
</tr>
<tr>
<td></td>
<td>Transmembrane domains prediction tools</td>
<td><a href="http://www.tripos.com/">http://www.tripos.com/</a></td>
<td>IRIX</td>
</tr>
<tr>
<td></td>
<td>TM-HMM</td>
<td><a href="http://www.cbs.dtu.dk/services/TMHMM-1.0/">http://www.cbs.dtu.dk/services/TMHMM-1.0/</a></td>
<td>On-line</td>
</tr>
<tr>
<td></td>
<td>TMPred</td>
<td><a href="http://www.ch.embnet.org/software/TMPRED_form.html">http://www.ch.embnet.org/software/TMPRED_form.html</a></td>
<td>On-line</td>
</tr>
<tr>
<td></td>
<td>TopPred 2</td>
<td><a href="http://bioweb.pasteur.fr/seqanal/interfaces/toppred.html">http://bioweb.pasteur.fr/seqanal/interfaces/toppred.html</a></td>
<td>On-line</td>
</tr>
<tr>
<td></td>
<td>HMMPERF</td>
<td><a href="http://www.enzim.hu/hmmtop/">http://www.enzim.hu/hmmtop/</a></td>
<td>On-line</td>
</tr>
</tbody>
</table>

### 3D modelling tools

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GenoMine</td>
<td><a href="http://www.bionformatics.uc.edu/genomine">http://www.bionformatics.uc.edu/genomine</a></td>
<td>Linux, Sun Solaris, Silicon Graphics</td>
</tr>
<tr>
<td></td>
<td>WHAT IF</td>
<td><a href="http://www.psimx.co.uk/products/naomi/">http://www.psimx.co.uk/products/naomi/</a></td>
<td>Silicon Graphics</td>
</tr>
<tr>
<td></td>
<td>Sybyl</td>
<td><a href="http://www.cmbi.kun.nl/1I00/WIVWWW/">http://www.cmbi.kun.nl/1I00/WIVWWW/</a></td>
<td>DEC-alpha, Sun, IBM PC running Linux or DOS</td>
</tr>
<tr>
<td></td>
<td>Transmembrane domains prediction tools</td>
<td><a href="http://www.tripos.com/">http://www.tripos.com/</a></td>
<td>IRIX</td>
</tr>
<tr>
<td></td>
<td>TM-HMM</td>
<td><a href="http://www.cbs.dtu.dk/services/TMHMM-1.0/">http://www.cbs.dtu.dk/services/TMHMM-1.0/</a></td>
<td>On-line</td>
</tr>
<tr>
<td></td>
<td>TMPred</td>
<td><a href="http://www.ch.embnet.org/software/TMPRED_form.html">http://www.ch.embnet.org/software/TMPRED_form.html</a></td>
<td>On-line</td>
</tr>
<tr>
<td></td>
<td>TopPred 2</td>
<td><a href="http://bioweb.pasteur.fr/seqanal/interfaces/toppred.html">http://bioweb.pasteur.fr/seqanal/interfaces/toppred.html</a></td>
<td>On-line</td>
</tr>
<tr>
<td></td>
<td>HMMPERF</td>
<td><a href="http://www.enzim.hu/hmmtop/">http://www.enzim.hu/hmmtop/</a></td>
<td>On-line</td>
</tr>
</tbody>
</table>

### 3D protein modelling

Changes in protein sequence might alter the protein or non-protein ligand binding site and general conformation of the protein. Such changes are best analysed on a 3D model. The main limitation of this approach is the need for a crystal structure of the protein. So far very few models of proteins expressed in the eye exist. Although there are ways of building models of unknown proteins, the methods are still under development and no good solutions exist as yet. At the moment in silico mutational analysis can be performed on the existing models using any of the current software tools (Table 3). It is possible to generate insertions, deletions or amino acid substitutions and analyse their potential effects on ligand binding (protein–protein interaction). Molecular modelling also provides a way of assaying potential therapeutic agents. As we can expect an increase in the number of crystal structures available, the knowledge of how to use these structures will become increasingly important.
USING POLYMORPHIC VARIANTS AND KNOWN MUTATIONS FOR ASSOCIATION STUDIES AND SCREENING CLINICAL SAMPLES

There are two types of screening of clinical samples that are performed in studies of inherited eye diseases. Firstly, screening of large populations using polymorphic markers for association studies of multigenic diseases and secondly, screening of particular genes for mutations in pedigrees.

In the case of multigenic diseases such as age-related macular degeneration (AMD), which affects 50 per cent of the population over 65 years, screening of large patient and control populations is required to establish a correlation between an allele variant and a disease. In this case polymorphism databases such as the SNP division of GenBank or any of the others also reviewed recently are very useful in identifying potential variants in known genes. These can be used for identification of genes for screening and designing the screening tests to identify the presence of particular polymorphisms. A recent paper by Sunyaev et al. showed that 3D analysis can be combined with SNP data to maximise the information provided by SNP databases.

In the case of single gene disorders the ability to rapidly determine which part of the gene of interest contains mutational hotspots is important for screening large genes (over 20 kb). For this reason the databases of mutations in genes expressed in the eye (Table 4) are particularly useful as is a general human mutation database HGMD. The specialised databases are rapidly developing and are currently holding information on over 20 genes involved in ocular diseases (eg X-linked retinoschisis, vitelliform macular dystrophy, retinoblastoma). While access to some databases is restricted to registered users (such as Mutview based in Japan) most are freely accessible. The information provided varies with each database but generally both clinical and mutation information are provided.

TREATMENT STRATEGIES

As mentioned above, identification of model animals by database searching can help in the development of a gene or pharmacological therapy. Even more important in the design of drugs targeting specific enzymes or signalling pathways is the information on the protein structure and function that can be predicted or uncovered by database mining. The most useful tool in the near future will be 3D analysis of the effect of pharmacological agents on the structure and function of the protein.

CONCLUSIONS AND PERSPECTIVES

One of the aims of the Human Genome Project was to provide rapidly accessible information about genes in normal and disease states. As human genome sequencing draws to a conclusion the genes responsible for various diseases will be much easier to identify. It is, however, imperative that people involved in

Table 4: Eye disease mutation databases

<table>
<thead>
<tr>
<th>Disease/gene</th>
<th>Database location</th>
</tr>
</thead>
<tbody>
<tr>
<td>MutView – 19 genes</td>
<td><a href="http://mutview.dmb.med.keio.ac.jp">http://mutview.dmb.med.keio.ac.jp</a></td>
</tr>
<tr>
<td>Peripherin/RDS and rhodopsin</td>
<td><a href="http://mol.ophth.uiowa.edu/MOL_WWW/Compendium.html">http://mol.ophth.uiowa.edu/MOL_WWW/Compendium.html</a></td>
</tr>
<tr>
<td>Vitelliform macular dystrophy</td>
<td><a href="http://www.uni-wuerzburg.de/humangenetics/vmd2.html">http://www.uni-wuerzburg.de/humangenetics/vmd2.html</a></td>
</tr>
<tr>
<td>X-linked retinoschisis</td>
<td><a href="http://www.dmd.nl/rs/rshome.html">http://www.dmd.nl/rs/rshome.html</a></td>
</tr>
<tr>
<td>Retinoblastoma</td>
<td><a href="http://www.d-lohmann.de/Rb/mutations.html">http://www.d-lohmann.de/Rb/mutations.html</a></td>
</tr>
</tbody>
</table>

Only a fraction of eye diseases are listed in dedicated mutation databases
researching the inherited eye diseases are aware of how to extract the information from the available databases and how to analyse the mutations detected in patients’ samples. There are a number of reasons why more effort should be directed towards the use, development and promotion of bioinformatic tools for research into hereditary eye diseases:

- An immense amount of sequence data exist in databases and should be more fully utilised to find good candidate genes faster than traditional methods.

- An increase in the number of expression databases and specific disease databases would be advantageous. This would enable determination of whether a gene is expressed in the eye and if there are any diseases already associated with it.

- Use of computational methods can help to re-classify retinal and macular diseases based on underlying genetic causes. Correlation of genotype/protein function with phenotype will facilitate accurate diagnosis of patients and aid in development of therapeutic approaches.

- Difficulty in obtaining human tissue. Computational methods that can analyse a protein before any laboratory analysis is performed are very attractive as they can help in formulating a working hypothesis and identify a good animal model in which the function of a gene can be studied.

The two main areas that will help in further work on the inherited eye diseases will be, firstly, development of an integrated database incorporating the computational tools for protein analysis and, secondly, improvements in 3D modelling, which will enable analysis of protein–protein interactions and the influence of mutations on protein folding and structure. We hope that this review highlights the potential of bioinformatics to accelerate and augment traditional approaches to the analysis of inherited eye diseases.

References


6. URL: http://www.sph.uth.tmc.edu/Retnet/


11. URL: http://www.ensembl.org


17. URL: http://www.hgnp.mrc.ac.uk


24. URL: http://genome.sgc.riken.go.jp


27. URL: http://telethon.bio.unipd.it/GETMaps


32. URL: www.informatics.jax.org/


38. URL: http://www.gcg.com


43. URL: http://jura.ebi.ac.uk:8888/


46. URL: http://www.sgi.com/chembio/resources/clustalw/


The study of eye diseases


76. URL: http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_prosite.html
77. URL: http://www.sanger.ac.uk/software/pfam/search.shtml
83. URL: http://archive.uwcm.ac.uk/uwcm/ng/hgmd/.
84. URL: http://mutview.dnb.med.keio.ac.jp