D148–D152 *Nucleic Acids Research, 2004, Vol. 32, Database issue*
DOI: 10.1093/nar/gkh090

EyeSite: a semi-automated database of protein families in the eye

David A. Lee[1,2,*], Sandrine Fefeu[1], Adrian A. Edo-Ukeh[1,2], Christine A. Orengo[2] and Christine Slingsby[1]

[1]Department of Crystallography, Birkbeck College, University of London, Malet Street, London WC1E 7HX, UK and [2]Biomolecular Structure and Modelling Group, Department of Biochemistry and Molecular Biology, University College London, Gower Street, London WC1E 6BT, UK

Received August 15, 2003; Revised and Accepted October 7, 2003

ABSTRACT

The EyeSite is a web-based database of protein families for proteins that function in the eye and their homologous sequences. The resource clusters proteins at different levels of homology in order to faciltate functional annotation of sequences and modelling of proteins from structural homologues. Eye proteins are organized into the tissue types in which they function and are clustered into homologous families using a novel protocol employing the TribeMCL algorithm. Homologous families are further subdivided into sequence clusters for which multiple sequence alignments are generated. Structural annotations from the CATH domain database are provided for nearly 90% of the sequences, and protein family annotations from the Pfam database for ~86%. Homology models have also been generated where appropriate. The EyeSite is stored in a relational database and is extensively linked to other online bioinformatics resources to help relate allelic variants, annotations and clinical details to the derived data in the database. The EyeSite is available for online search, sequence information and model retrieval at http://eyesite.cryst.bbk.ac.uk/.

INTRODUCTION

Knowledge of the cellular origin, structure and function of all the proteins expressed in ocular tissues is required to fully understand eye biology and its diseases. Bioinformatic methods can contribute significantly to the acquisition and presentation of this knowledge. Databases representing the transcript profile of the specialized tissues of the eye assembled by Dr Graeme Wistow of the National Eye Institute (1) are contributing new information about which proteins function in the eye, and computational tools are required to help interpret these data. The development of the EyeSite database is centred around protein family relationships and their linkage to protein structure. The focus of this first release of the database is on the mainly tissue-specific eye proteins from many species organized with their homologues into families and given structural annotations. As such, the scope of the database is significantly different from other available resources such as the KmeyeDB database (2) of mutations in the human eye, which does not include homologues and makes no links to structural data.

The experimentally solved 3D structures of proteins that are currently publicly available are described and classified in CATH (3) and SCOP (4). These databases decompose proteins into structural domains defined as distinct, compact and stable protein structural units that fold independently of each other and which typically act as independent functional and evolutionary units. The sequences of these domains may be used to construct profiles with which to predict the domain architectures of proteins which possess regions of similar sequence. Gene3D (5) is a database of CATH-derived structural annotations for complete genomes and is now approaching an average coverage of 50% of the sequences annotated within any genome. Pfam (6) also provides a comprehensive and high-quality set of profiles for protein domain families. Pfam domains are functional domains that sometimes correspond to structural domains but which may also span several structural domains or have no identifiable structure. As such CATH and Pfam domain predictions complement each other very advantageously and both are presented in the EyeSite. Homology models are built for those sequences that are sufficiently close to a solved structure to be reliable (7).

Function is a multifaceted attribute and insight can be gained by organizing proteins into homologous families. Protein sequences in the database are derived from text and sequence searches of GenBank (8) and are clustered into homologous families that correlate well with biological function and domain architecture using the TribeMCL algorithm (9), described below. Hand edited, generalized functional descriptions are provided for each family. Subsequent sub-clustering of the families on the basis of sequence identity allows the sequence neighbourhood of each protein to be investigated at several different levels of similarity (e.g. 35%, 60%, 95% and 100% sequence identity) and for biologically meaningful sequence alignments to be constructed. All protein sequences can be retrieved in FASTA format files for easy import into other sequence analysis programs.

*To whom correspondence should be addressed. Tel: +44 20 7679 3890; Fax: +44 20 7679 7193; Email: dlee@biochem.ucl.ac.uk

Swiss-Prot (10) provides high-quality functional annotations for many of the proteins that are contained in GenBank. LocusLink (11) provides access to diverse information on human diseases including, for example, genetic loci, phenotypes, MIM entries and clinical data. These data are amalgamated in a relational database with the new derived data generated for the EyeSite. The procedures used to create the EyeSite are shown in Figure 1a. In addition to keyword and accession code searches, facilities exist to browse the database, by tissue type, protein family and species.

CONSTRUCTION AND ORGANIZATION OF THE EyeSite

Sequence acquisition

Seed sequences are identified by submitting appropriate keywords to Entrez at GenBank (8) and then the results are manually filtered. Keywords used are: lens and eye; cornea and eye; trabecular meshwork; retina; optic and nerve and eye; iris and eye; fovea and eye; ciliary body; choroid not plexus. This results in the identification of a representative set of 3000 'seed' sequences. The sequence search tool PSI-BLAST (12) is then used for the detection of homologues within the GenBank NRDB100. Because PSI-BLAST is very computationally expensive, non-redundant representatives are first selected from the seed sequence set. All-against-all sequence alignments are performed using a standard Needleman–Wunsch dynamic programming algorithm, Homol (3) and then single linkage clustering is performed at several levels of sequence identity using Seqcluster (3). Representatives are chosen from all clusters at 35% sequence identity and used as query sequences in the PSI-BLAST search. The threshold E-value for matches is 0.1 and each match has to align with at least 50% of the query sequence. These parameters are deliberately non-stringent since further filtering of the matches is carried out at later stages in the protocol. The 50% length overlap cut-off prevents many of the small protein fragments in GenBank from being selected. Protein accession numbers are used to map to Swiss-Prot, taxon, LocusLink and MIM numbers.

Sequence clustering

The TribeMCL method developed by Enright *et al.* (9) relies on the Markov cluster (MCL) algorithm for the assignment of proteins into families based on pre-computed sequence similarity information. In essence the method identifies 'natural' clusters in a protein–protein similarity graph in a process that is sensitive to the density and strength of connections. A similarity matrix is generated from an all-against-all comparison of the sequences using gapped BLAST (12) with a threshold E-value for matches of 0.0001 as suggested by Enright (personal communication). Only those clusters that contain at least one seed eye sequence are retained. Homologous families created in the TribeMCL clustering are further clustered according to sequence identity using Homol and Seqcluster, as above, to produce clusters at the 35% (S35), 60% (S60), 95% (S95) and 100% (S100) levels. This produces progressively tighter functional and structural groupings and makes clear the range of variation across a homologous family. Generalized functional names are assigned to each validated homologous family through visual inspection of their members.

Sequence alignment

ClustalW (13) multiple alignments are generated for all of the EyeSite families (with <1000 members) and clusters. The user may 'zoom in and out' on sequence alignments to identify conserved/variable positions at different levels of clustering, which can assist greatly in the analysis of mutation data and of sequence–function relationships. Redundancy is removed from the alignment displayed for the chosen family/cluster by selecting one representative from each cluster in the next level down of the clustering hierarchy. The user may select an individual sequence to act as a representative.

CATH and Pfam domain architecture prediction

Distant structural predictions are made by screening sequences against SAM-T99 (14) generated Hidden Markov Models (HMMs) for the CATH structural domain representatives. The fw0.7 script as recommended in the documentation is used for all SAM model building. Because screening against these models is very computationally expensive and was developed for the detection of very remote structural similarities, only representative structures at the S35 level are used. Matches are processed using a modified version of the DomainFinder method (15), which filters out insignificant matches on the basis of a combination of E-value and overlap in the sequence alignment and then attempts to identify structural domain boundaries from the remaining matches.

All EyeSite sequences are also screened against SAM HMMs generated from the Pfam-A alignments downloaded from the Pfam FTP site. Matches are accepted using the same cut-off parameters as are used in the DomainFinder method since here we are searching for homologues rather than attempting to make definitive functional assignments. Domain architecture diagrams show the regions that match CATH and Pfam profiles so the two sources of annotations may be compared.

Homology modelling

Because a relatively sparse set of representative CATH domain profiles are used for structure prediction some close relationships can be overlooked. BLAST searches of all PDB (16) sequences are performed to identify exact or close structural matches. Exact structural matches are defined as one or more experimentally solved structures completely matching over the full sequence and close structural matches are defined as aligning with at least 80% of the EyeSite sequence and having at least 60% sequence identity with it. Sequences with close structural matches are used as targets for homology modelling in a high-throughput automatic procedure which produces a Needleman–Wunsch sequence alignment with the template of highest sequence identity and then models the structure of the target using Modeller 6v2 developed by the Šali group (17).

Database schema

All data are loaded into a MySQL database according to a schema developed by Shepherd *et al*. (18).

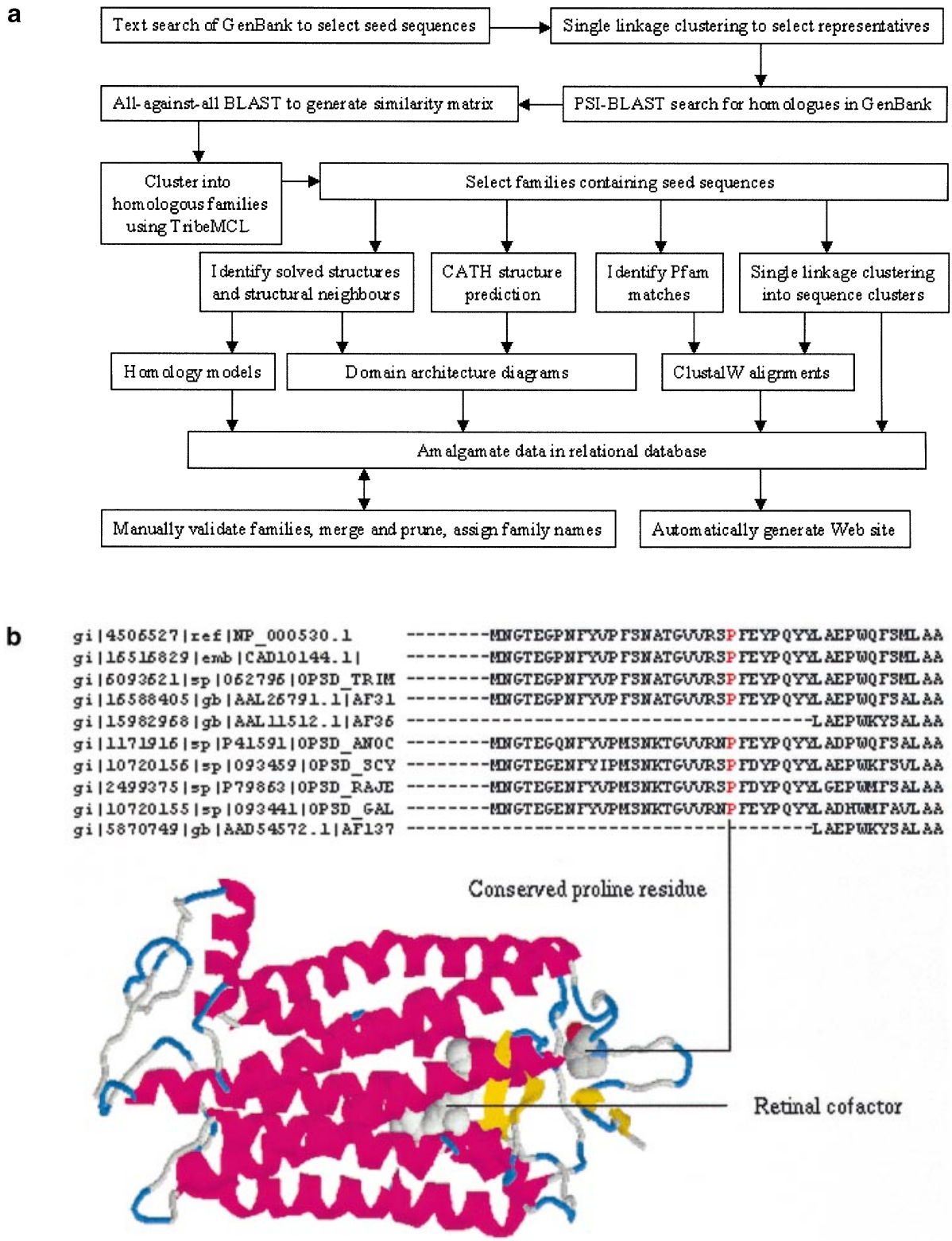

Figure 1. (**a**) Flowchart of the pocedure used to construct the EyeSite. (**b**) Location of the mutated proline residue (P23H) in human rhodopsin, from retinal family 1 of G-protein-coupled receptors, subfamily visual opsins, which is associated with retinitis pigmentosa (MIM entry 180380). An excerpt of the S60 level ClustalW alignment is shown indicating the conservation of this residue in those sequences that contain this N-terminal segment. Human rhodopsin is in the top line of the excerpt. The position of this proline residue is indicated in a RasMol cartoon representation of the homology model of human rhodopsin. The probable position of the retinal cofactor is also indicated.

Table 1. Number of clusters in the EyeSite at the tissue type, homologous family (H) level, and at the 35% (S35), 60% (S60), 95% (S95) and 100% (S100) levels of clustering by sequence identity

Level of clustering	Number of clusters
Tissue type	9
H	722
S35	7124
S60	25730
S95	49285
S100	75665

CONTENTS OF CURRENT RELEASE

Several thousand sequences of proteins known to function in nine ocular tissues have been grouped into 722 families. When their homologues are included this expands to 76 119 protein sequences in this first release of the EyeSite. The grouping of homologous families and sequence clusters is summarized in Table 1. Sequence alignments are provided for all homologous families (with <1000 members) and all sequence clusters.

A total of 794 (1%) of the EyeSite sequences have exact structural matches and a further 11 280 (14.8%) have close structural matches from which homology models are constructed. A further 55 793 (73.3%) of sequences have structural annotations for at least one structural domain, so that in total 89.1% of the EyeSite sequences have a CATH structural annotation. In addition 65 380 (85.9%) of the EyeSite sequences have a Pfam-A annotation. Only 2166 (2.8%) of the EyeSite sequences have neither a CATH nor a Pfam-A annotation while 59 294 (77.9%) have both types of annotation. All sequences and models may be retrieved and links may be followed to the predicted CATH and Pfam families. Links may also be followed to Swiss-Prot annotations at EMBL and to the GenBank entries for the proteins and their taxon, LocusLink and MIM entries.

DATABASE ACCESS

The EyeSite is entered through an initial web page (http://eyesite.cryst.bbk.ac.uk/) which allows the user to either browse the database or make specific queries. All subsequent pages are generated 'on the fly' using SQL queries of the underlying MySQL database executed from within Perl CGI scripts. Lists of generalized functional names for homologous families associated with particular tissue types may be accessed through the clickable eye diagram. A list for the complete eye is also provided. Alternatively, the 'Keyword Search' or 'Search by Code' boxes may be used to search by free text, GI code, Swiss-Prot, LocusLink or MIM identifier. Free text searches may be restricted to selected species for the whole database or be restricted to within a family. Each homologous family name provides a link to a hierarchical arrangement of pages for each sequence cluster within the family. The hierarchy may be descended using the cluster level and number links. Initially only representatives for each sub-cluster are displayed but a tool bar at the top of each page allows all sequences in that cluster to be viewed. The tool bar may also be used to select the display of sequence alignments and domain architectures and to download all sequences in the current cluster in FASTA format. Domain architecture pages provide links to predicted CATH and Pfam families and to homology models, which may be either downloaded in PDB coordinate format or displayed using RasMol. Each individual protein entry consists of an annotation derived from Swiss-Prot where available or else the GenBank annotation. Links are provided to GenBank and, when available, to SwissProt, LocusLink and MIM. A mapping between a sequence alignment and the location of a disease associated mutation in an homology model is shown in Figure 1b.

CONCLUSIONS AND PERSPECTIVES

The EyeSite has been developed both to organize proteins that function in the eye into homologous families and sequence clusters and to help investigate the relationship between protein structure and the underlying causes of disease. The protein families and their CATH and Pfam annotations are a rich source of information for functional and evolutionary studies and the alignments of the sequence clusters facilitate the study of important conserved amino acids, mutations and polymorphisms and the construction of sequence profiles. We have endeavoured to provide integrated access to both clinical details and protein structure information in order to promote their association.

ACKNOWLEDGEMENTS

We thank Dr Graeme Wistow for inspiration and useful discussions, Dr Anton Enright for the TribeMCL method and Drs David Houldershaw and Duncan McKenzie for systems administration. The financial support of the Medical Research Council, London, is gratefully acknowledged.

REFERENCES

1. Wistow,G. (2002) A project for ocular bioinformatics: NEIBank. *Mol. Vis.*, **8**, 161–163.
2. Minoshima,S., Mitsuyama,S., Ohno,S., Kawamura,T. and Shimizu,N. (2000) Eye Disorder Database 'KmeyeDB'. *Hum. Mutat.*, **15**, 95–98.
3. Pearl,F.M., Bennett,C.F., Bray,J.E., Harrison,A.P., Martin,N., Shepherd,A., Sillitoe,I., Thornton,J. and Orengo,C.A. (2003) The CATH database: an extended protein family resource for structural and functional genomics. *Nucleic Acids Res.*, **31**, 452–455.
4. LoConte,L., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2002) SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res.*, **30**, 264–267.
5. Buchan,D.W., Shepherd,A.J., Lee,D., Pearl,F.M., Rison,S.C., Thornton,J.M. and Orengo,C.A. (2002) Gene3D: structural assignments for whole genes and genomes using the CATH domain structure database. *Genome Res.*, **12**, 503–14.
6. Bateman,A., Birney,E., Cerruti,L., Durbin,R., Etwiller,L., Eddy,S.R., Griffiths-Jones,S., Howe,K.L., Marshall,M. and Sonnhammer,E.L. L. (2002) The Pfam Protein Families Database. *Nucleic Acids Res.*, **30**, 276–280.
7. Vitkup,D., Melamud,E., Moult,J. and Sander,C. (2001) Completeness in structural genomics. *Nature Struct. Biol.*, **8**, 559–566.
8. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J., Rapp,B.A. and Wheeler,L. (2000) GenBank. *Nucleic Acids Res.*, **28**, 15–18.
9. Enright,A.J., Van Dongen,S. and Ouzounis,C.A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, **30**, 1575–1584.
10. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.-C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan, I., Pilbout,S. and Schneider,M. (2003) The SWISS-PROT protein

knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
11. Pruitt,K.D. and Maglott,D.R. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.*, **29**, 137–140.
12. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402.
13. Higgins,D. Thompson,J., Gibson,T., Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTALW: improving the sensitivity of progressive multiple alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
14. Karplus,K., Barrett,C. and Hughey,R. (1998) Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, **14**, 846–856.
15. Pearl,F.M.G., Lee,D., Bray,J.E., Sillitoe,I., Todd,A.E., Harrison,A.P., Thornton,J.M. and Orengo,C.A. (2000) Assigning genomic sequences to CATH. *Nucleic Acids Res.*, **28**, 277–282.
16. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
17. Marti-Renom,M.A., Stuart,A., Fiser,A., Sánchez,R., Melo,F. and Šali,A. (2000) Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.*, **29**, 291–325.
18. Shepherd,A.J., Martin,N.J., Johnson,R.G., Kellam,P. and Orengo,C.A. (2002) PFDB: a generic protein family database integrating the CATH domain structure database with sequence based protein family resources. *Bioinformatics*, **18**, 1666–1672.