Published online 23 October 2008 *Nucleic Acids Research, 2009, Vol. 37, Database issue* ***D347–D354***
doi:10.1093/nar/gkn791

MODBASE, a database of annotated comparative protein structure models and associated resources

Ursula Pieper[1], Narayanan Eswar[1], Ben M. Webb[1], David Eramian[1,2], Libusha Kelly[1,3], David T. Barkan[1,3], Hannah Carter[4], Parminder Mankoo[4], Rachel Karchin[4], Marc A. Marti-Renom[5], Fred P. Davis[6] and Andrej Sali[1,*]

[1]Department of Bioengineering and Therapeutic Sciences, Department of Pharmaceutical Chemistry, and California Institute for Quantitative Biosciences, Byers Hall at Mission Bay, Office 503B, University of California at San Francisco, 1700 4th Street, San Francisco, CA 94158, [2]Graduate Group in Biophysics, [3]Graduate Group in Bioinformatics, University of California at San Francisco, CA, [4]Department of Biomedical Engineering, Institute for Computational Medicine, Johns Hopkins University, 3400 North Charles Street, Baltimore, MD 21218, USA, [5]Structural Genomics Unit, Bioinformatics & Genomics Department, Centro de Investigación Príncipe Felipe (CIPF), Avda. Autopista del Saler 16, Valencia 46012, Spain and [6]Howard Hughes Medical Institute, Janelia Farm, 19700 Helix Drive, Ashburn, VA 20147, USA

Received September 15, 2008; Accepted October 8, 2008

ABSTRACT

MODBASE (http://salilab.org/modbase) is a database of annotated comparative protein structure models. The models are calculated by MODPIPE, an automated modeling pipeline that relies primarily on MODELLER for fold assignment, sequence–structure alignment, model building and model assessment (http:/salilab.org/modeller). MODBASE currently contains 5 152 695 reliable models for domains in 1 593 209 unique protein sequences; only models based on statistically significant alignments and/or models assessed to have the correct fold are included. MODBASE also allows users to calculate comparative models on demand, through an interface to the MODWEB modeling server (http://salilab.org/modweb). Other resources integrated with MODBASE include databases of multiple protein structure alignments (DBAli), structurally defined ligand binding sites (LIGBASE), predicted ligand binding sites (AnnoLyze), structurally defined binary domain interfaces (PIBASE) and annotated single nucleotide polymorphisms and somatic mutations found in human proteins (LS-SNP, LS-Mut). MODBASE models are also available through the Protein Model Portal (http://www.proteinmodelportal.org/).

INTRODUCTION

The genome sequencing efforts are providing us with complete genetic blueprints for hundreds of organisms, including humans. We are now faced with the challenge of assigning, investigating and modifying the functions of proteins encoded by these genomes. This task is generally facilitated by 3D structures of the proteins (1–3), which are best determined by experimental methods such as X-ray crystallography and NMR-spectroscopy. The number of experimentally determined structures deposited in the Protein Data Bank (PDB) more than doubled from 23 096 to 52 821 over the last 5 years (September 2008) (4). However, the number of sequences in comprehensive sequence databases, such as UniProt (5) and GenPept (6), continues to grow even more rapidly than the number of known protein structures; for example, the number of sequences in UniProt increased from 1.2 million to 6.4 million over the same period. Therefore, protein structure prediction is essential for structural characterization of sequences without experimentally determined structures.

The most accurate models are generally obtained by homology or comparative modeling (7–10), which is applicable when an experimentally determined structure related to the target sequence is available. The fraction of sequences in a genome for which comparative models can be obtained automatically varies from ~20%–75% (11).

*To whom correspondence should be addressed. Tel: +1 415 514 4227; Fax: +1 415 514 4231; Email: sali@salilab.org

The process of comparative modeling usually requires the use of a number of programs to identify template structures, to generate sequence–structure alignments, to build the models and to evaluate them. In addition, various sequence and structure databases that are accessed by these programs are needed. Once an initial model is calculated, it is generally refined and ultimately analyzed in the context of many other related proteins and their functional annotations. Here, we describe MODBASE, a database of comparative protein structure models, and several associated databases and servers that facilitate modeling and analysis tasks for both expert and novice users. We highlight the improvements of MODBASE that were implemented since the last report (11), including updates in the modeling software, user interface and associated annotation tools. We also illustrate the utility of MODBASE by describing several projects depending on large model sets.

CONTENTS

Comparative modeling (MODELLER and MODPIPE)

Models in MODBASE are calculated using MODPIPE, our automated software pipeline for comparative modeling (12). It relies primarily on the various modules of MODELLER (13) for its functionality and is adapted for large-scale operation on a cluster of PCs using scripts written in PERL and Python. Sequence–structure matches are established using a variety of fold-assignment methods, including sequence–sequence (14), profile–sequence (15,16) and profile–profile alignments (16,17). Odds of finding a template structure are increased by using an *E*-value threshold of 1.0. By default, 10 models are calculated for each of the alignments (13). A representative model for each alignment is then chosen by ranking based on the atomic distance-dependent statistical potential DOPE (18). Finally, the fold of each model is evaluated using a composite model quality criterion that includes the coverage of the modeled sequence, sequence identity implied by the sequence–structure alignment, the fraction of gaps in the alignment, the compactness of the model and various statistical potential *Z*-scores (18–20). Only models that are assessed to have the correct fold were included in the final model sets.

A key feature of the pipeline is not prejudging the validity of sequence–structure relationships at the fold-assignment stage; instead, sequence–structure matches are assessed after the construction of the models and their evaluation. This approach enables a thorough exploration of fold assignments, sequence–structure alignments and conformations, with the aim of finding the model with the best evaluation score.

Comparative modeling web server (MODWEB)

MODWEB is our comparative modeling web server that is an integral module of MODBASE (http://salilab.org/modweb) (12). MODWEB accepts one or more sequences in the FASTA format and calculates their models using MODPIPE based on the best available templates from the PDB. Alternatively, MODWEB also accepts a protein structure as input, calculates a profile for each identifiable sequence homolog in the UniProt database, followed by modeling these homologs based on detectable templates in the PDB as well as the user-provided structure. Finally, MODWEB proposes a representative model based on model assessment. This module is a useful tool for measuring the impact of new structures, such as those generated by structural genomics efforts (21). The module allows us to assess the impact of a newly determined protein structure on the modeling of sequences of unknown structure. It is also used to identify new members of sequence superfamilies with at least one member of known structure. The results of MODWEB calculations are available to the users through the MODBASE interface as private datasets protected with passwords.

Pairwise and multiple structure alignments (DBAli)

DBAli (http://www.dbali.org/) stores pairwise comparisons of all structures in the PDB calculated using the program MAMMOTH (22), as well as multiple structure alignments generated by the SALIGN module of MODELLER-9 (23). DBAli contains approximately 1.7 billion pairwise comparisons and 12 732 family-based multiple structure alignments for 34 637 nonredundant protein chains out of 96 804 protein chains in the PDB. Additional information is provided by ModDom that assigns domain boundaries from structure and ModClus that allows the user to generate clusters of similar protein structures. These DBAli tools help users to analyze the protein structure space by establishing relationships between protein structures and their fragments in a flexible and dynamic manner.

Ligand binding sites (LIGBASE and AnnoLyze)

The LIGBASE module stores a list of the binding sites of known structure for approximately 230 000 ligands found in the PDB (24). The ligands include small molecules, such as metal ions, nucleotides, saccharides and peptides. Binding sites in all known structures are defined to consist of residues with at least one atom within 5 Å of any ligand atom. For each template structure, MODBASE also contains a list of putative binding sites that were predicted by the AnnoLyze program (25). The predictions are based on inheriting an actual binding site from any related known structure if at least 75% of the binding site residues are within 4 Å of the template residues in a global superposition of the two structures in DBALI and if at least 75% of the binding site residue types are invariant. In addition, the putative ligand binding sites in the models are then mapped *via* the target–template alignments. The putative ligand binding sites are stored as SITE records and the binding site membership frequency per residue is indicated in the B-factor column of the model coordinate files. Sixty-five percent of MODBASE models have at least one predicted binding site.

Protein interactions (PIBASE)

PIBASE (http://pibase.janelia.org, http://salilab.org/pibase) is a comprehensive database of structurally defined protein interfaces (26). It is composed of binary interfaces

between pairs of chains or domains extracted from structures in the PDB and the Probable Quaternary Structure server PQS using domain assignments from the Structural Classification of Proteins and CATH fold classification systems. PIBASE currently contains 269 821 SCOP, 269 438 CATH, and 216 739 chain binary interfaces. A diverse set of geometrical, physiochemical and topological properties are calculated for each complex, its domains, interfaces and binding sites. The database is accessible through the web server and can also be installed locally. The software used to build PIBASE is available for download under an open-source license.

PIBASE is a convenient resource for structural information on protein–protein interactions and is easily integrated with other databases. It is currently used by the AnnoLyze annotation program (27) and the LS-SNP annotation system (28). The complexes stored in PIBASE can also be used as templates to predict the composition and structure of protein complexes using comparative modeling followed by an assessment of the modeled interface (29). This approach was applied to predict host–pathogen interactions for 10 'neglected' human pathogens (30).

Single nucleotide polymorphisms and somatic mutations (LS-SNP and LS-Mut)

LS-SNP [http://karchinlab.org/LS-SNP, http://salilab.org/LS-SNP (28)] and LS-Mut [http://karchinlab.org/LS-Mut, (31,32)] are collections of annotated DNA sequence variants in protein-coding exons that result in an amino acid residue-type substitution. These resources focus on inherited genetic variants and tumor-derived somatic mutations, respectively. For LS-SNP, genomic locations of the variants are taken from the dbSNP database (33) and are mapped onto as many human proteins in the UniProt database (34) as possible. The mapping is achieved *via* a collection of protein-to-mRNA and mRNA-to-genome alignments produced with the Known Genes algorithm (35). For LS-Mut, somatic mutation data from tumor sequencing projects are used, consisting of transcript identifiers from RefSeq, CCDS and Ensembl (36,37), codon positions and amino acid residue-type substitutions. Our software then maps the mutations onto translated protein sequences. LS-Mut currently includes mutations from 24 advanced pancreatic cancers and 22 glioblastoma multiforme (brain) tumors. For both LS-SNP and LS-Mut, human protein sequences are aligned with homologous proteins of known structure from PDB, to build comparative protein structure models using MODPIPE. Models are constructed for all significant alignments covering a distinct region of protein sequence (*E*-value cutoff 0.0001). UCSF Chimera (38) is used to visualize the location of the residue substitutions on the model. We use our software and DSSP (39) to identify secondary structure elements and relative solvent accessibility of the residue positions. Putative protein and small ligand binding sites on the models are annotated with PIBASE and the LIGBASE module of MODBASE, respectively, to infer which SNPs or somatic mutations may destabilize protein quaternary structure or interfere with small molecule ligand binding.

MODBASE MODEL SETS

Models in MODBASE are organized into a number of datasets. The largest dataset contains models of all sequences in the UniProt database that are detectably related to at least one known structure in the PDB from July 2005. Because of the rapid growth of the public sequence databases, we now concentrate our efforts on adding datasets that are useful for specific projects, rather than attempt to model all known protein sequences with detectable template structures. Currently, MODBASE includes datasets of nine archaeal genomes, 13 bacterial genomes and 18 eukaryotic genomes (Table 1). Together with other project-oriented datasets, MODBASE currently contains 5 152 695 models from domains in 1 593 209 unique sequences. Next, we illustrate the utility of MODBASE by outlining several recent projects.

Structural genomics of the enolase and amidohydrolase superfamilies

Comparative models of enzymes in the amidohydrolase and enolase superfamilies have contributed to studying their substrate specificity by the Enzyme Specificity Consortium (ENSPEC) as well as selecting targets for a structural genomics effort by the New York SGX Research Center for Structural Genomics (NYSGXRC). In particular, we selected 535 target proteins from 130 genomes for high-throughput structure determination by X-ray crystallography, resulting in 61 unique structures thus far. Both template-based modeling and sequence-based modeling were essential in identifying suitable targets.

Structural genomics of membrane proteins

Comparative modeling was also applied to inform target selection for the structural genomics of membrane proteins as part of the Center for Structures of Membrane Proteins (CSMP) at UCSF (40). The goal of CSMP is to express, purify and determine the structures of representative members of integral membrane protein classes. MODBASE models were combined with an interactive web-based target selection tool to facilitate selection of biologically interesting targets with little or no structural data available. In addition, template-based modeling in MODWEB is being used to calculate how many sequences can be modeled based on newly determined CSMP structures.

ABC Transporters

ABC transporters are a large and diverse set of integral membrane proteins that couple the action of ATP binding, hydrolysis and release to substrate transport across a cellular membrane (41). Mutations in 13 of the 48 human ABC transporters are associated with monogenic human disease phenotypes (42). Additional variants are being

Table 1. MODBASE datasets

Dataset/Project	Taxonomy ID	No. of Transcripts	No. of Sequences modeled	No. of Models	Sequence source
Genomes (*genomes for the TDI)					
Archaea					
Archaeoglobus fulgidus	2234	2409	1794	3980	NCBI
Methanococcus jannaschii	2190	1785	1480	1707	NCBI
Nanoarchaeum equitans	160 232	536	447	496	NCBI
Picrophilus torridus	82 076	1535	1260	2902	NCBI
Pyrobaculum aerophilum	13 773	2600	1566	3497	NCBI
Pyrococcus furiosus	2261	2113	1524	3373	NCBI
Sulfolobus solfataricus	2287	2922	2006	4451	NCBI
Thermoplasma volcanium	50 339	1497	1204	2806	NCBI
Thermoplasma acidophilum		1480	1220	2801	NCBI
Bacteria					
Bacillus subtilis	1423	4105	3374	9245	NCBI
Burkholderia mallei	13 373	4798	3910	23 219	NCBI
Clostridium tetani	1513	2413	2158	5864	NCBI
Escherichia coli	562	4206	3150	5994	NCBI
*Mycobacterium leprae**	1769	1605	1178	2493	OrthoMCL-DB
*Mycobacterium tuberculosis**	1773	3991	2808	5913	TubercuList
Mycoplasma pneumoniae	2104	687	426	857	NCBI
Pseudomonas aeruginosa	287	5559	3806	9222	NCBI
Rickettsia prowazekii	782	835	754	2136	NCBI
Staphylococcus aureus MRSA252	282 458	2635	1184	3161	NCBI
Streptococcus pyogenes	1314	1691	1440	3984	NCBI
*Wolbachia**	953	805	621	1873	TIGR
Yersinia pestis	632	3882	3215	8371	NCBI
Eukaryota					
Arabidopsis thaliana	3702	30 707	23 807	70 494	ENSEMBL
*Brugia malayi**	6279	11 397	7850	23 219	TIGR
Caenorhabditis elegans	6239	22 698	18 996	52 235	NCBI
Canis familiaris	9615	30 264	22 614	65 617	ENSEMBL
*Cryptosporidium hominis**	237 895	3886	1614	3287	CryptoDB
*Cryptosporidium parvum**	5807	3806	1918	3969	CryptoDB
Danio rerio	Calculation in progress				ENSEMBL
Drosophila melanogaster	7227	17 104	9381	24 683	NCBI
*H.sapiens**	9606	32 010	21 270	51 084	OrthoMCL-DB
*Leishmania major**	5664	8274	3975	8285	GeneDB
Mus musculus	10 090	30 133	25 338	70 783	NCBI
Pan troglodytes	Calculation in progress				ENSEMBL
*Plasmodium falciparum**	5833	5363	2599	5053	PlasmoDB
*Plasmodium vivax**	5855	5342	2359	4670	PlasmoDB
Rattus norvegicus	Calculation in progress				ENSEMBL
Saccharomyces cerevisiae	4932	6600	3035	5543	NCBI
*Schistosoma mansoni**	6183	25 304	8576	26 076	GeneDB
*Toxoplasma gondii**	5811	7793	1530	3064	ToxoDB
*Trypanosoma brucei**	5691	9210	3900	8054	GeneDB
*Trypanosoma cruzi**	5693	19 607	7390	14 858	GeneDB
Xenopus laevis	8355	27 952	25 457	69 191	NCBI
Selected projects					
CSMP datasets		195 235	184 139	690 255	GENPEPT NR
NYSGXRC datasets		553 537	493 672	1 415 237	GENPEPT NR
Enzyme Specificity Project		15 833	10 875	183 591	SFLD/NR
ABC Transporter		152	85	85	
GPCR		11 586	11 551	24 272	
UNIPROT Datasets 2005		1 742 816	1 025 196	2 146 830	UNIPROT
Total (including other datasets)		2 608 987	1 593 209	5 152 695	

The sequences were retrieved from ENSEMBL (36), TIGR (50), NCBI-Genbank (6), OrthoMCL-DB (51), TubercuList (52), CryptoDB (53), GeneDB (54), ToxoDB (55), SFLD (56) and UniProt (34).

identified in hundreds of individuals by the Pharmacogenomics of Membrane Transporters (PMT) consortium at UCSF (43). To annotate these variants, we modeled nucleotide binding and membrane spanning domains with detectably related template structures in all human ABC transporters. The dataset also includes models of sequences with disease-associated and polymorphic non-synonymous SNPs found in the nucleotide binding domains. Finally, the incomplete or unsatisfactory modeling coverage was used to suggest specific targets for a structural genomics effort on ABC transporters by CSMP.

Human caspases

Caspases are cysteine proteases involved in multiple apoptotic pathways. An experimental approach was recently developed to identify caspase substrates by biotinylating natural protein N-termini and selecting protein fragments containing unblocked α-amines characteristically generated upon proteolytic cleavage (44). Likely high accuracy models of protein substrates prior to cleavage were identified in the MODBASE human genome datasets and analysis of the structural properties of the cleavage sites was performed. While these sites often appeared in disordered, solvent accessible regions of the substrate as expected (45), a surprising number were found in α-helices and partially inaccessible regions, information which can now be incorporated into new algorithms for predicting additional caspase substrates.

Binding sites and ligands for the tropical disease initiative

Open source drug discovery is an alternative avenue to conventional patent-based drug development, illustrated by the proposed Tropical Disease Initiative (TDI) (http://tropicaldisease.org) (46). Open source drug discovery involves a decentralized, web-based and community-wide collaboration, in which scientists from laboratories, universities, institutes and corporations volunteer to work together for a common cause. To contribute to this effort, we calculated comparative protein structure models for 10 genomes of organisms that cause 'neglected' tropical diseases (Table 1). We followed up by predicting binding sites for known drugs using the AnnoLyze program (25). These predictions may be used as a starting point for experimentally testing the biological functions of the target proteins and potentially even as leads for drug discovery.

Host–pathogen protein interactions for TDI

Pathogens have evolved numerous strategies to infect their hosts, while hosts have evolved immune responses and other defenses to these foreign challenges. The vast majority of host–pathogen interactions involve protein–protein recognition, yet our current understanding of these interactions is limited. We developed and applied a computational whole-genome protocol that generates testable predictions of host–pathogen protein interactions (30) (http://salilab.org/hostpathogen). The protocol first scans the host and pathogen genomes for proteins with similarity to known protein complexes, then assesses these putative interactions, using structure if available, and, finally, filters the remaining interactions using biological context, such as the stage-specific expression of pathogen proteins and tissue expression of host proteins. The technique was applied to 10 pathogens, using their MODBASE model datasets. Several specific predictions have been made that warrant experimental follow-up, including interactions from previously characterized mechanisms, such as cytoadhesion and protease inhibition, as well as suspected interactions in hypothesized networks, such as apoptotic pathways.

G-Protein Coupled receptors

G-protein coupled receptors (GPCR) are a large family of pharmacologically important transmembrane receptors that are involved in the recognition of a wide variety of extra-cellular ligands. It has been estimated that this family of proteins is the target for about half of all currently marketed drugs. Atomic structures are known for only three sub-families of GPCRs, including light-sensitive rhodopsins, β1 and β2 adrenergic receptors that all belong to the Class A Rhodopsin-like family (GPCRDB nomenclature). The GPCR dataset in MODBASE consists of models for approximately 12 000 UniProt sequences that are related to one of these structures. The models span several sub-families of the Class A Rhodopsin-like family, including aminergic, peptide, hormone, opsin, olfactory and nucleotide receptors. These models are used for ligand docking and virtual screening computations by DOCK (47).

ACCESS AND INTERFACE

The main access to MODBASE is through its web interface at http://salilab.org/modbase, by querying with Uniprot and GI identifiers, gene names, annotation keywords, PDB codes, datasets, organisms, sequence similarity to the modeled sequences (BLAST) and model-specific criteria such as model reliability, model size and target–template sequence identity. Additionally, it is possible to retrieve coordinate files, alignment files and ligand-binding information in text files. Select genome datasets are also available from our ftp server (ftp://salilab.org/databases/modbase/projects).

The output of a search is displayed on pages with varying amounts of information about the modeled sequences, template structures, alignments and functional annotations. An example of the output from a search resulting in one model is shown in Figure 1. A ribbon diagram of the model with the highest target–template sequence identity is displayed by default, together with details of the modeling calculation. Ribbon thumbprints of additional models for this sequence link to corresponding pages with more information. The ribbon diagrams are generated on the fly using Molscript (48) and Raster3D (49). A pull-down menu provides links to additional functionality: the ligand-binding module, the SNP module, retrieval of coordinate and alignment files, as well as molecular visualization by Chimera that allows the user to display template and model coordinates together with their alignment. If mutation information is available for a protein sequence, links to the details are provided in the cross-references section. Additionally, cross-references to various other databases, including PDB, UniProt, SwissProt/TrEMBL, PubMed and the UCSC Genome Browser, are given. Other MODBASE pages provide overviews of more than one sequence or structure. All MODBASE pages are interconnected to facilitate easy navigation between different views.

Access through external databases

MODBASE models in academic and public datasets are directly accessible from several other databases, including

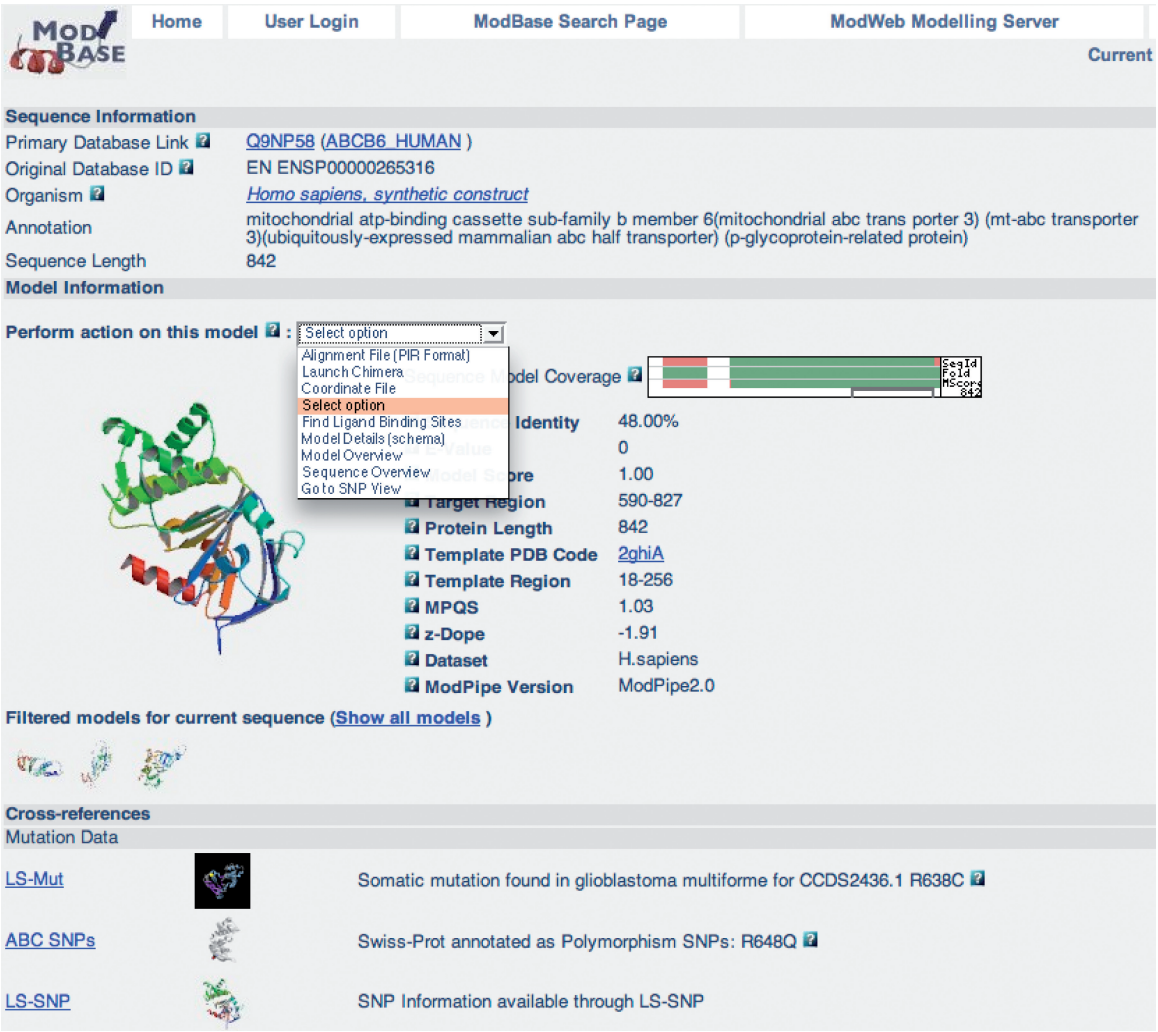

Figure 1. MODBASE Model Details page (Example Q9NP58 from the human genome dataset): this page provides links to all models for this specific sequence. A ribbon diagram of the primary model, database annotations and modeling details are displayed. Links to additional models for different target regions or models from other datasets are displayed as thumbprints. The pull-down menu provides access to alternative MODBASE views and other types of information (if available), such as data about mutations and putative ligand binding sites. The cross-references section contains links to relevant internal and external databases. For this particular sequence, mutation data are available from LS-Mut, LS-SNP and ABC SNPs.

the SwissProt/TrEMBL sequence pages, UniProt, PIR's iProClass, EBI's InterPro, the UCSC Genome Browser and PubMed (LinkOut). Importantly, MODBASE models are also accessible through the Protein Model Portal (http://proteinmodelportal.org), a module of the Protein Structure Initiative Knowledgebase (PSI KB). The Model Portal has the potential to become the single entry point for users interested in experimentally determined or computationally predicted models. For a user query, the portal will interrogate participating source model databases and modeling servers to provide a comprehensive view of all available models of the query sequence.

FUTURE DIRECTIONS

MODBASE will grow by adding models calculated on demand by external users (using MODWEB) as well as our own calculations of model datasets that are needed for our research projects (using MODPIPE, MODWEB or MODELLER). These updates will reflect improvements in the methods and software used for calculating the models as well as the new template structures in the PDB and new sequences in UniProt. In the future, we expect that most of the users will access MODBASE models through the Protein Model Portal.

CITATION

Users of MODBASE are requested to cite this article in their publications.

ACKNOWLEDGEMENTS

We are grateful to Tom Ferrin, Daniel Greenblatt, Conrad Huang and Tom Goddard for CHIMERA and

contributing to the MODBASE/CHIMERA interface. For linking to MODBASE from their databases, we thank Torsten Schwede (Protein Model Portal), David Haussler and Jim Kent (UCSC Genome Browser), Amos Bairoch (SwissProt/TrEMBL), Rolf Apweiler (InterPro), Patsy Babbitt (SFLD) and Cathy Wu (PIR/iProClass). We are also grateful for computing hardware gifts from Mike Homer, Ron Conway, NetApp, IBM, Hewlett Packard and Intel.

FUNDING

National Institutes of Health (R01 GM54762, U54 GM074945, U54 GM074929, U01 GM61390, P01 GM71790 to A.S., GM08284 to D.E., NSF EF 0626651); the Sandler Family Supporting Foundation (to A.S.); Susan G. Komen Foundation (KG080137 to R.K.); Spanish Ministerio de Educación y Ciencia (BIO2007/66670 to M.A.M-R). Funding for open access charge: U54 GM074945.

REFERENCES

1. Domingues,F.S., Koppensteiner,W.A. and Sippl,M.J. (2000) The role of protein structure in genomics. *FEBS Lett.*, **476**, 98–102.
2. Brenner,S.E. and Levitt,M. (2000) Expectations from structural genomics. *Protein Sci.*, **9**, 197–200.
3. Skolnick,J., Fetrow,J.S. and Kolinski,A. (2000) Structural genomics and its importance for gene function analysis. *Nat. Biotechnol.*, **18**, 283–287.
4. Deshpande,N., Addess,K.J., Bluhm,W.F., Merino-Ott,J.C., Townsend-Merino,W., Zhang,Q., Knezevich,C., Xie,L., Chen,L., Feng,Z. *et al.* (2005) The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema. *Nucleic Acids Res.*, **33**, D233–D237.
5. Bairoch,A., Apweiler,R., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al.* (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **33**, D154–D159.
6. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2008) GenBank. *Nucleic Acids Res.*, **36**, D25–D30.
7. Baker,D. and Sali,A. (2001) Protein structure prediction and structural genomics. *Science*, **294**, 93–96.
8. Wallner,B. and Elofsson,A. (2005) All are not equal: a benchmark of different homology modeling programs. *Protein Sci.*, **14**, 1315–1327.
9. Hillisch,A., Pineda,L.F. and Hilgenfeld,R. (2004) Utility of homology models in the drug discovery process. *Drug Discov. Today*, **9**, 659–669.
10. Eswar,N., Webb,B., Marti-Renom,M.A., Madhusudhan,M.S., Eramian,D., Shen,M.Y., Pieper,U. and Sali,A. (2007) Comparative protein structure modeling using MODELLER. *Curr. Protocols Protein Sci./editorial board, John E. Coligan . . . et al.*, **Chapter 2**, Unit 29.
11. Pieper,U., Eswar,N., Davis,F.P., Braberg,H., Madhusudhan,M.S., Rossi,A., Marti-Renom,M., Karchin,R., Webb,B.M., Eramian,D. *et al.* (2006) MODBASE: a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res.*, **34**, D291–D295.
12. Eswar,N., John,B., Mirkovic,N., Fiser,A., Ilyin,V.A., Pieper,U., Stuart,A.C., Marti-Renom,M.A., Madhusudhan,M.S., Yerkovich,B. *et al.* (2003) Tools for comparative protein structure modeling and analysis. *Nucleic Acids Res.*, **31**, 3375–3380.
13. Sali,A. and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, **234**, 779–815.
14. Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.
15. 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.
16. Eswar,N., Webb,B., Marti-Renom,M.A., Madhusudhan,M.S., Eramian,D., Shen,M.Y., Pieper,U. and Sali,A. (2006) Comparative protein structure modeling using Modeller. *Curr. Protocols Bioinformatics/editoral board, Andreas D. Baxevanis . . . et al.*, **Chapter 5**, Unit 56.
17. Marti-Renom,M.A., Madhusudhan,M.S. and Sali,A. (2004) Alignment of protein sequences by their profiles. *Protein Sci.*, **13**, 1071–1087.
18. Shen,M.Y. and Sali,A. (2006) Statistical potential for assessment and prediction of protein structures. *Protein Sci.*, **15**, 2507–2524.
19. Eramian,D., Shen,M.Y., Devos,D., Melo,F., Sali,A. and Marti-Renom,M.A. (2006) A composite score for predicting errors in protein structure models. *Protein Sci.*, **15**, 1653–1666.
20. Melo,F., Sanchez,R. and Sali,A. (2002) Statistical potentials for fold assessment. *Protein Sci.*, **11**, 430–448.
21. Chance,M.R., Fiser,A., Sali,A., Pieper,U., Eswar,N., Xu,G., Fajardo,J.E., Radhakannan,T. and Marinkovic,N. (2004) High-throughput computational and experimental techniques in structural genomics. *Genome Res.*, **14**, 2145–2154.
22. Ortiz,A.R., Strauss,C.E. and Olmea,O. (2002) MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci.*, **11**, 2606–2621.
23. Marti-Renom,M.A., Ilyin,V.A. and Sali,A. (2001) DBAli: a database of protein structure alignments. *Bioinformatics*, **17**, 746–747.
24. Stuart,A.C., Ilyin,V.A. and Sali,A. (2002) LigBase: a database of families of aligned ligand binding sites in known protein sequences and structures. *Bioinformatics*, **18**, 200–201.
25. Marti-Renom,M.A., Rossi,A., Al-Shahrour,F., Davis,F.P., Pieper,U., Dopazo,J. and Sali,A. (2007) The AnnoLite and AnnoLyze programs for comparative annotation of protein structures. *BMC Bioinformatics*, **8(Suppl. 4)**, S4.
26. Davis,F.P. and Sali,A. (2005) PIBASE: a comprehensive database of structurally defined protein interfaces. *Bioinformatics*, **21**, 1901–1907.
27. Marti-Renom,M.A., Pieper,U., Madhusudhan,M.S., Rossi,A., Eswar,N., Davis,F.P., Al-Shahrour,F., Dopazo,J. and Sali,A. (2007) DBAli tools: mining the protein structure space. *Nucleic Acids Res.*, **35**, D393–D397.
28. Karchin,R., Diekhans,M., Kelly,L., Thomas,D.J., Pieper,U., Eswar,N., Haussler,D. and Sali,A. (2005) LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources. *Bioinformatics*, **21**, 2814–2820.
29. Davis,F.P., Braberg,H., Shen,M.Y., Pieper,U., Sali,A. and Madhusudhan,M.S. (2006) Protein complex compositions predicted by structural similarity. *Nucleic Acids Res.*, **34**, 2943–2952.
30. Davis,F.P., Barkan,D.T., Eswar,N., McKerrow,J.H. and Sali,A. (2007) Host pathogen protein interactions predicted by comparative modeling. *Protein Sci.*, **16**, 2585–2596.
31. Jones,S., Zhang,X., Parsons,D.W., Lin,J.C., Leary,R.J., Angenendt,P., Mankoo,P., Carter,H., Kamiyama,H., Jimeno,A. *et al.* (2008) Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science*, **321**, 1801–1806.
32. Parsons,D.W., Jones,S., Zhang,X., Lin,J.C., Leary,R.J., Angenendt,P., Mankoo,P., Carter,H., Siu,I.M., Gallia,G.L. *et al.* (2008) An integrated genomic analysis of human Glioblastoma multiforme. *Science*, **321**, 1807–1812.
33. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
34. Wu,C.H., Apweiler,R., Bairoch,A., Natale,D.A., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R. *et al.* (2006) *Nucleic Acids Res.*, **34**, D187–191.
35. Hsu,F., Kent,W.J., Clawson,H., Kuhn,R.M., Diekhans,M. and Haussler,D. (2006) The UCSC known genes. *Bioinformatics*, **22**, 1036–1046.
36. Flicek,P., Aken,B.L., Beal,K., Ballester,B., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cunningham,F., Cutts,T. *et al.* (2008) Ensembl 2008. *Nucleic Acids Res.*, **36**, D707–D714.

37. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S. *et al.* (2008) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **36**, D13–D21.
38. Pettersen,E.F., Goddard,T.D., Huang,C.C., Couch,G.S., Greenblatt,D.M., Meng,E.C. and Ferrin,T.E. (2004) UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.*, **25**, 1605–1612.
39. Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
40. Li,M., Hays,F.A., Roe-Zurz,Z., Vuong,L., Kelly,L., Robbins,R., Ho,C.M., Pieper,U., O'Connell,J., Miercke,L.J. *et al.* (2008) Eukaryotic Integral Membrane Protein Production For Structural Genomics. *J. Mol. Biol., in press.*
41. Dean,M., Rzhetsky,A. and Allikmets,R. (2001) The human ATP-binding cassette (ABC) transporter superfamily. *Genome Res.*, **11**, 1156–1166.
42. Hamosh,A., Scott,A.F., Amberger,J.S., Bocchini,C.A. and McKusick,V.A. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **33**, D514–D517.
43. Leabman,M.K., Huang,C.C., DeYoung,J., Carlson,E.J., Taylor,T.R., de la Cruz,M., Johns,S.J., Stryke,D., Kawamoto,M., Urban,T.J. *et al.* (2003) Natural variation in human membrane transporter genes reveals evolutionary and functional constraints. *Proc. Natl Acad. Sci. USA*, **100**, 5896–5901.
44. Mahrus,S., Trinidad,J.C., Barkan,D.T., Sali,A., Burlingame,A.L. and Wells,J.A. (2008) Global sequencing of proteolytic cleavage sites in apoptosis by specific labeling of protein N termini. *Cell*, **134**, 866–876.
45. Hubbard,S.J., Campbell,S.F. and Thornton,J.M. (1991) Molecular recognition. Conformational analysis of limited proteolytic sites and serine proteinase protein inhibitors. *J. Mol. Biol.*, **220**, 507–530.
46. Maurer,S.M., Rai,A. and Sali,A. (2004) Finding cures for tropical diseases: is open source an answer? *PLoS Med.*, **1**, e56.
47. Hermann,J.C., Marti-Arbona,R., Fedorov,A.A., Fedorov,E., Almo,S.C., Shoichet,B.K. and Raushel,F.M. (2007) Structure-based activity prediction for an enzyme of unknown function. *Nature*, **448**, 775–779.
48. Kraulis,P.J. (1991) MOLSCRIPT: a program to produce both detailed and schematic plorts of protein structures. *J. Appl. Crystallogr.*, **24**, 946–950.
49. Merritt,E.A. and Bacon,D.J. (1997) Raster3D: photorealistic molecular graphics. *Methods Enzymol.*, **277**, 505–524.
50. Ghedin,E., Wang,S., Spiro,D., Caler,E., Zhao,Q., Crabtree,J., Allen,J.E., Delcher,A.L., Guiliano,D.B., Miranda-Saavedra,D. *et al.* (2007) Draft genome of the filarial nematode parasite Brugia malayi. *Science*, **317**, 1756–1760.
51. Chen,F., Mackey,A.J., Stoeckert,C.J. Jr. and Roos,D.S. (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res.*, **34**, D363–D368.
52. Cole,S.T. (1999) Learning from the genome sequence of Mycobacterium tuberculosis H37Rv. *FEBS Lett.*, **452**, 7–10.
53. Heiges,M., Wang,H., Robinson,E., Aurrecoechea,C., Gao,X., Kaluskar,N., Rhodes,P., Wang,S., He,C.Z., Su,Y. *et al.* (2006) CryptoDB: a Cryptosporidium bioinformatics resource update. *Nucleic Acids Res.*, **34**, D419–D422.
54. Hertz-Fowler,C., Peacock,C.S., Wood,V., Aslett,M., Kerhornou,A., Mooney,P., Tivey,A., Berriman,M., Hall,N., Rutherford,K. *et al.* (2004) GeneDB: a resource for prokaryotic and eukaryotic organisms. *Nucleic Acids Res.*, **32**, D339–D343.
55. Gajria,B., Bahl,A., Brestelli,J., Dommer,J., Fischer,S., Gao,X., Heiges,M., Iodice,J., Kissinger,J.C., Mackey,A.J. *et al.* (2008) ToxoDB: an integrated Toxoplasma gondii database resource. *Nucleic Acids Res.*, **36**, D553–D556.
56. Pegg,S.C., Brown,S.D., Ojha,S., Seffernick,J., Meng,E.C., Morris,J.H., Chang,P.J., Huang,C.C., Ferrin,T.E. and Babbitt,P.C. (2006) Leveraging enzyme structure-function relationships for functional inference and experimental design: the structure-function linkage database. *Biochemistry*, **45**, 2545–2555.