2014

Genome3D

BBSRC
bioscience for the future

[WORKSHOP 2014]

“How to Annotate Your Sequence Using Structure” – an introduction to Genome3D and associated Structural Bioinformatics resources.

Contents

Workshop Information

Date:	Thursday 24th July 2014
Location:	Room 1.02, Malet Place Engineering Building, UCL
Time:	10:00 - 16:30

Schedule

10:00 - 10:30	**Arrival, welcome and coffee**
10:30 - 10:40	**Opening remarks by Christine Orengo**
10:40 - 11:25	**Invited talk from Torsten Schwede** *"Modeling protein structures and complexes using evolutionary information"*
11:25 - 12:10	**Invited talk from Geoff Barton** *"Jalview, what it is, what it does and why it is important!"*
12:10 - 13:20	**Lunch (please make own arrangements)**
13:20 - 14:40	**Talks by Genome3D partners:** Tom Blundell, Alexey Murzin, Julian Gough, Mike Sternberg, Christine Orengo, David Jones
14:40 - 15:00	**Talk on Genome3D**
15:00 - 15:30	**Transfer to computer cluster room + coffee break**
15:30 - 16:30	**Tutorial**

Lunch

Please make your own arrangements for lunch during the 70 minute lunch break. There are many places near UCL from which you can buy lunch, for example:

- **Tottenham Court Road and nearby**:
  Pret a Manger, Marks and Spencer, Eat, Sainsbury's, Tesco, Tapped and Packed
- **Torrington Place:**
  Planet Organic, Costa in Waterstones
- **University shops/canteens**:
  UCL refectory, UCL Shop, Birkbeck cafe (with Costa)

Tutorials

http://genome3d.eu

(click “Tutorials” at top of the page)

Keynote Talks

Professor Torsten Schwede
Biozentrum, University of Basel

"Modeling protein structures and complexes using evolutionary information"

ABSTRACT: Protein structure homology modelling has become a key technique for exploring sequence-structure-function relationships by extrapolating the available experimental structure information to structurally uncharacterized protein sequences. In recent years, these techniques have matured into reliable automated processes, which allow homology models to be used in a wide spectrum of applications in life science research. Quality estimation is a crucial step in structure prediction, as ultimately the quality of a model determines its usefulness for specific biomedical applications. In this presentation, I will introduce modern tools for automated homology modelling and discuss strengths and limitations of current approaches.

Professor Geoff Barton
Division of Computational Biology, University of Dundee

"Jalview, what it is, what it does and why it is important!"

ABSTRACT: Jalview is a free, open-source, multiple sequence alignment and analysis workbench that considers sequences from a structural perspective. It supports alignment by 8 popular multiple alignment algorithms, interactive editing, integration of remote web services (e.g. secondary structure and disorder prediction) and integration of 3D structure analysis through Jmol. It also supports tools for RNA alignment and secondary structure prediction. This talk will give an overview of Jalview's features and explain future development directions.

Genome3D: Primary Resources

An overview of the primary contributing resources within the Genome3D database.

Professor Tom L. Blundell

Department of Biochemistry, University of Cambridge, UK

TOCCATA & FUGUE - http://structure.bioc.cam.ac.uk/toccata

Managed by Dr Bernardo Ochoa

FUGUE is a remote homology detection program whose approach is characterised by two features: the use of Environment-Specific Substitution Tables (ESSTs) in structural profiles and an automatic alignment algorithm selection with structure-dependent gap penalties.

Residues in protein structures exist in various structural environments, such as secondary structure, solvent exposure or hydrogen bonding, which, in combination, can affect the conservation of residues in evolution, as well as the probability of deletions and insertions. By using BLOSUM-like substitution tables and algorithms that take these facts into consideration, FUGUE is able to detect homology that might be missed by less sophisticated methods.

FUGUE relies on the TOCCATA database for domain assignments from both SCOP and CATH domain classifications, taking over from the original HOMSTRAD. TOCCATA creates single and multi-domain consensus profiles from SCOP families and CATH superfamilies and annotates PDBs according to their conformational state, such as ligand binding and oligomeric state.

The TOCCATA website allows searching a query sequence against its database of profiles and the alignment of a sequence against any profile, as well as further refinement with a customised profile from a selection of templates.

TOCCATA & FUGUE form the foundation of the VIVACE pipeline, which uses the output to generate multi-template models using MODELLER and annotating them with XSuLT, described below. It also implements a variety of quality assessment metrics to annotate models according to their reliability.

Shi J, Blundell TL and Mizuguchi K. FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. Journal of Molecular Biology 310, no. 1 (June 29, 2001): 243–57. PMID: 11419950

Professor Julian Gough

Computational Genomics, Bristol University, UK

SUPERFAMILY - http://www.supfam.org

Managed By Prof Julian Gough

SUPERFAMILY is a database of structural and functional annotation for all proteins and genomes.

The SUPERFAMILY annotation is based on a collection of hidden Markov models, which represent structural protein domains at the SCOP superfamily level. A superfamily groups together domains which have an evolutionary relationship. The annotation is produced by scanning protein sequences from over 3200 completely sequenced genomes against the hidden Markov models.

Using superfamily you can submit peptide sequences for SCOP classification, view phylogenetic tree and domain assignments for each genome, view functional annotation, Gene Ontology annotation, InterPro abstract and genome assignments for each superfamily and explore taxonomic distribution of a superfamily across the tree of life.

SUPERFAMILY is a member of the InterPro consortium of protein annotation databases, and has been integrated into the Ensembl eukaryotic genome project and The Arabidopsis Information Resource. To date, the SUPERFAMILY publications have been cited over 1,000 times. SUPERFAMILY has been used in structural, functional, evolutionary and phylogenetic research projects.

Gough, J., Karplus, K., Hughey, R. and Chothia, C. (2001). "Assignment of Homology to Genome Sequences using a Library of Hidden Markov Models that Represent all Proteins of Known Structure." J. Mol. Biol., 313(4), 903-919.

Professor Mike Sternberg

Centre for Integrative Systems Biology and Bioinformatics, Imperial College London

Phyre2 - http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index

Lead developer: Lawrence Kelley

Phyre2 is a server for predicting protein structure from sequence. Phyre2 uses the alignment of hidden Markov models via HHsearch to significantly improve accuracy of alignment and detection rate. Phyre2 incorporates a new *ab initio* folding simulation called Poing to model regions of your proteins with no detectable homology to known structures. Poing is also used to combine multiple templates. Distance constraints from individual models are treated as linear elastic springs. Poing then synthesises your entire protein in the presence of these springs and at the same time models unconstrained regions using its physics simulation. Phyre2 also provides predicted secondary structure, transmembrane and disordered regions using tool from Prof David Jones's group. Phyre2 has a series of extra features including

- **Phyre Alarm** which if you have not found a good confident hit you can add your sequence to your phyre alarm account and it will be automatically scanned against new entries to our fold library every week and you will be notified of any new confident hits.
- **One to one threading** which is used if you have a protein sequence and you want to model it on a specific template of your choosing.
- **BackPhyre** which instead of predicting the 3D structure of a protein sequence, helps users who have a solved structure and they are interested in determining if there is a related structure in a genome of interest.
- **Phyre Investigator** which provides extensive analyses of the predicted structures including using SuSPect to predict the phenotypic effect of a sequence variant.

Professor Christine Orengo

Institute of Structure and Molecular Biology, University College London, UK

CATH - http://www.cathdb.info

Managed by Dr Ian Sillitoe

CATH (Class, Architecture, Topology, Homology) is a semi-automatic, classification of protein domains. (C)lass refers to the composition of alpha-helices and beta strands. (A)rchitecture describes the orientation of the secondary structures in 3D whilst (T)opology captures both orientation and connectivity of these elements. The most important level in the classification is the (H)omologous superfamily level which groups together domains which are related through evolution. Structural domains in CATH are extracted from proteins deposited in the PDB. This is done using algorithms that look for structural similarity with a domain already in CATH or recognise novel domains. All new domains with no close homologues in CATH are manually inspected to obtain accurate domain boundaries. To classify new domains, their structure is scanned against representatives from each fold group and homologous superfamily in CATH, using the SSAP and CATHEDRAL algorithms. Sequence searches are also performed against sequence patterns (HMMs) for each superfamily. Domains having similar folds to existing CATH domains are assigned to the same fold group. If they also have significant sequence similarity to domains in a superfamily in that fold group, or clear functional similarity, the domain is classified in that superfamily. If not, it seeds a new superfamily. If the new domain has no significant matches with any CATH domain it is manually inspected to determine whether it has one of the existing CATH architectures or a novel architecture.

CATH version 4.0 contains 277,687 domains, 2,738 homologous superfamilies and 1,375 fold groups. Within each superfamily we have classified relatives into functional families (FunFams) if they share sequence patterns reflecting significant similarity in functional properties. Our CATH sequence domain search facility provides FunFam annotations for all domains matching the query. The superfamily pages in CATH also provide a range of other information on - the multidomain contexts in which domains in the family are found; phylogenetic information on when functions emerged during evolution (via the FunTree resource); structural variation across the superfamily.

Sillitoe I, Cuff AL, Dessailly BH, Dawson NL, Furnham N, Lee D, Lees JG, Lewis TE, Studer RA, Rentzsch R, Yeats C, Thornton JM, Orengo CA. New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures. Nucleic Acids Res. 2013 Jan;41(Database issue):D490-8.

GENE3D - http://gene3d.biochem.ucl.ac.uk

Managed by Dr Jon Lees

Gene3D is a database of protein domain structure annotations for protein sequences. Domains are predicted using a library of profile HMMs from 2,738 CATH superfamilies. Gene3D assigns domain annotations to Ensembl and UniProt sequence sets including >6000 cellular genomes and >20 million unique protein sequences. The coverage is expanded by integrating Pfam and SUPERFAMILY domain annotations, and resolving domain overlaps to provide highly comprehensive composite multi-domain architectures. These data are made accessible for comparative genome analyses through novel search algorithms for searching genomes to identify related multi-domain architectures. Gene3D is a member of InterPro.

Lees JG, Lee D, Studer RA, Dawson NL, Sillitoe I, Das S, Yeats C, Dessailly BH, Rentzsch R, Orengo CA. Gene3D: Multi-domain annotations for protein sequence and comparative genome analysis. Nucleic Acids Res. 2014 42(Database issue):D240-5.

Professor David Jones

Department of Computer Science, University College London, UK

pDomTHREADER - http://bioinf.cs.ucl.ac.uk/psipred/?pdomthreader=1

Managed by Dr Federico Minneci

pDomTHREADER is a reliable and sensitive tool for structural domain superfamily discrimination, which is available through the PSIPRED Protein Analysis Workbench at the above URL. It combines information from both sequence and structure to produce highly accurate domain alignments. In particular, the program first builds a PSSM (Position Specific Scoring Matrix, or profile) for the input sequence, and then aligns it against a template library of representative structural domains using secondary-structure dependent gap-penalties and classic pair- and solvation potentials.

Lobley A, Sadowski MI, Jones DT. pGenTHREADER and pDomTHREADER: new methods for improved protein fold recognition and superfamily discrimination. Bioinformatics. 2009 Jul 15;25(14):1761-7

DomSerf – http://bioinf.cs.ucl.ac.uk/psipred/?domserf=1

Managed by Dr Federico Minneci

DomSerf is a fully automated homology modelling pipeline. It integrates common tools for protein fold recognition and 3D structure prediction of individual protein domains. DomSerf initially collects suitable templates by running both PSI-BLAST and pDomTHREADER searches against appropriate libraries using conservative cut-offs, so that both close and distant evolutionary relationships to domains of know structure can be identified. These hits are screened and assembled into the domain architecture with maximum coverage of the input sequences and with minimum overlap between domain assignments at the same time using DomainFinder. Finally, the resulting set of domains and their alignments to the target sequences are used to generate the output three dimensional structure predictions with MODELLER.

Lewis TE, Sillitoe I, Andreeva A, Blundell TL, Buchan DW, Chothia C, Cuff A, Dana JM, Filippis I, Gough J, Hunter S, Jones DT, Kelley LA, Kleywegt GJ, Minneci F, Mitchell A, Murzin AG, Ochoa-Montaño B, Rackham OJ, Smith J, Sternberg MJ, Velankar S, Yeats C, Orengo C. Genome3D: a UK collaborative project to annotate genomic sequences with predicted 3D structures based on SCOP and CATH domains. Nucleic Acids Res. 2013 Jan;41(Database issue):D499-507

Genome3D: Additional Resources

An overview of some additional, freely available resources maintained by Genome3D contributors.

Professor Tom L. Blundell

Department of Biochemistry, University of Cambridge, UK

CHOPIN - http://structure.bioc.cam.ac.uk/chopin

Managed by Dr Bernardo Ochoa

CHOPIN is a database and web resource of homology models of *Mycobacterium tuberculosis* (strain H37Rv) built with the full capabilities of the VIVACE pipeline. Each target with significant FUGUE hits has been modelled in a variety of potential conformations, according to the state of available templates on the TOCCATA database. The database includes 5,268 hits for 2,911 of the 4,008 proteins and 13,169 different models.

Additional features an automatically updated list of all published experimental structures and stability predictions of a list of mutations potentially related to antibiotic resistance.

CREDO - http://structure.bioc.cam.ac.uk/credo

Created by Dr Adrian Schreyer, maintained by Dr Bernardo Ochoa

CREDO is a structural interactomics database for drug discovery. It consists of a relational database storing all pairwise atomic interactions of inter- as well as intra-molecular contacts between small- and macromolecules found in experimentally determined structures from the PDB, along with programming and web APIs for accessing the information.

It provides a variety of cheminformatics functions to query chemical components present in the database, and implements a mapping to structural variations provided by EnsEMBL Variation, allowing the linking of phenotypes to binding sites or protein-protein interfaces.

Schreyer AM, Blundell TL. CREDO: a structural interactomics database for drug discovery. Database. 2013 Jul 18;2013(0):bat049.

mCSM / DUET - http://structure.bioc.cam.ac.uk/mcsm | duet

Created by Dr Douglas Pires

mCSM is a method to predict the effects of missense mutations on proteins. mCSM is a machine-learning method that uses signatures representing the geometric and physicochemical environment of a residue to predict the effects of point mutations on protein stability, as well as their effects on protein-protein and protein-nucleic acid affinities.

DUET combines mCSM with the previously developed, complementary approach SDM to provide a consensus prediction, using support vector machines to provide an optimized merged predictor, improving the accuracy compared to either method individually.

The mCSM and DUET web servers allow users to upload protein structures and run predictions for mutations of interest.

Pires DEV, Ascher DB, Blundell TL. mCSM: predicting the effects of mutations in proteins using graph-based signatures. Bioinformatics. 30.3 (2014): 335-342.

Pires DEV, Ascher DB, Blundell TL. DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. Nucleic Acids Research. 42.W1 (2014): W314-W319.

Professor Julian Gough

Computational Genomics, Bristol University, UK

Proteome Quality Index - http://pqi-list.org

A database attempting to provide a measure of proteome quality available from a comprehensive database of downloadable proteomes.

dcGO - http://supfam.org/SUPERFAMILY/dcGO

dcGO integrates knowledge from a variety of contexts, ranging from functional information like Gene Ontology (GO) to others on enzymes and pathways, from phenotype information across major model organisms to information about human diseases and drugs.

Database of Disordered Protein Predictions - http://d2p2.pro

Pre-computed disorder predictions on a large library of proteins from completely-sequenced genomes. Goals of the database include making statistical comparisons of the various prediction methods freely available to the prediction community, as well as facilitating biological investigation of the disordered protein space.

More information on other projects from the group:

http://bioinformatics.bristol.ac.uk/software.php

Professor Mike Sternberg

Centre for Integrative Systems Biology and Bioinformatics, Imperial College London

SuSPect - http://www.sbg.bio.ic.ac.uk/suspect/about.html

Lead developer: Chris Yates

SuSpect is a server for predicting the phenotypic effects of non-synonymous single nucleotide polymorphisms (nsSNPs). By integrating sequence, structural and systems biology-based features, SuSPect predicts how likely an nsSNP is to cause disease and gives an explanation of how it may have its effects. Imput consists of protein sequence data, a protein structure or a VCF file.

Yates, C.M., I. Filippis, L.A. Kelley & M.J. Sternberg, SuSPect: Enhanced Prediction of Single Amino Acid Variant (SAV Phenotype Using Network Features. J Mol Biol, 2014. 426: p. 2692-701.

CombFunc - http://www.sbg.bio.ic.ac.uk/~mwass/combfunc/

Lead developer: Mark Wass

CombFunc is an automated method for the prediction of protein function. Users can either submit a sequence or use a uniprot identifier. CombFunc runs multiple analyses for the query sequence (details below) to obtain data that can be associated with protein function. This data is then combined using a machine learning approach (Support Vector Machines SVM) resulting in a function prediction.
Wass, M.N., G. Barton & M.J. Sternberg, *CombFunc: predicting protein function using heterogeneous data sources.* Nucleic Acids Res, 2012. **40**: p. W466-70.

3DLigandSite - http://www.sbg.bio.ic.ac.uk/3dligandsite/

Lead developer: Mark Wass

3DLigandSite predicts ligand binding sites using ligand binding data from homologous structures. Users can either input a sequence, or a protein structure. Where a sequence is used, Phyre, our in-house structure prediction server first models the protein structure.

*Wass, M.N., L.A. Kelley & M.J.E. Sternberg, 3DLigandSite: predicting ligand-binding sites using similar structures. Nucleic Acid Res, 2010. **38**: p. W469 – W437.*

WINARNS - http://www.sbg.bio.ic.ac.uk/~winarns/

Lead developer: Hang Phan

WINARNS is a server for global alignment of protein-protein interaction networks (PPINs) using random neural network (RNN) formulation. The WINARNS web server provides both precaculated alignments for different pairs of species (see Precalculated Alignments) and a service to align user-generated protein interaction networks using our web server (fill in form below to submit alignment jobs).

Phan, H.T., M.J.E. Sternberg & E. Gelenbe, Aligning protein-protein interaction networks using random neural networks, in 2012 IEEE International Conference on Bioinformatics and Biomedicine2012: Philadelphia, PA, USA

Professor Christine Orengo

Institute of Structure and Molecular Biology, University College London, UK

CATHEDRAL Server – http://www.cathdb.info/search/by_structure

The CATHEDRAL server can be used to search CATH superfamilies with a query protein structure. It returns a list of domain structures which match all or part of the query structure together with superpositions of the structures, Information on the structure similarity, number of residues aligned and sequence identity.

Redfern OC, Harrison A, Dallman T, Pearl FM, Orengo CA. CATHEDRAL: a fast and effective algorithm to predict folds and domain boundaries from multidomain protein structures. PLoS Comput Biol. 2007 Nov;3(11):e232

FUN-L - http://funl.org

Managed by Dr Jon Lees

Fun-L (Functional Lists) is a tool for target prioritisation for experimentalists. Given a set of query genes known to be involved in a pathway of interest, the remainder of the genome is ranked by likelihood of shared pathway membership to this initial query.

Most biological processes remain only partially characterised with many components still to be identified. Given that a whole genome can usually not be tested in a functional assay, identifying genes most likely to be of interest is of critical importance to avoid wasting resources. Given a set of known functionally-related genes and using a state-of-the-art approach to data integration and mining, our FUN-L (Functional Lists) method provides a ranked list of candidate genes for testing. Validation of predictions from FUN-L with independent RNAi screens confirms that FUN-L-produced lists are enriched in genes with the expected phenotypes.

Hériché JK, Lees JG, Morilla I, Walter T, Petrova B, Julia Roberti M, Hossain MJ, Adler P, Fernández JM, Krallinger M, Haering CH, Vilo J, Valencia A, Ranea JA, Orengo C, Ellenberg J. Integration of biological data by kernels on graph nodes allows prediction of new genes involved in mitotic chromosome condensation. Mol Biol Cell. 2014 Jun 18. pii: mbc.E13-04-0221.

PAIN Networks - http://www.PainNetworks.org

Managed by Dr Jon Lees

The Pain Networks resource integrates interaction data from various public databases with information on known pain genes and allows the user to examine a gene (or set of genes) of interest alongside known interaction partners. This information is displayed by the resource in the form of a network.

Hundreds of genes are proposed to contribute to nociception and pain perception. Historically, most studies of pain-related genes have examined them in isolation or alongside a handful of other genes. More recently the use of systems biology techniques has enabled us to study genes in the context of the biological pathways and networks in which they operate. The user can enrich these networks by using data from pain-focused gene expression studies to highlight genes that change expression in a given experiment or pairs of genes showing correlated expression patterns across different experiments. Genes in the networks are annotated in several ways including biological function and drug binding.

Perkins JR, Lees J, Antunes-Martins A, Diboun I, McMahon SB, Bennett DL, Orengo C. PainNetworks: a web-based resource for the visualisation of pain-related genes in the context of their network associations. Pain. 2013 Dec;154(12):2586.e1-12. Epub 2013 Sep 11. Review.

Professor David Jones

Department of Computer Science, University College London, UK

PSIPRED Protein Analysis Workbench - http://bioinf.cs.ucl.ac.uk/psipred/

Managed by Dr Federico Minneci

The PSIPRED Protein Analysis Workbench has a user-friendly interface that allows to launch at the same time several protein structural and functional annotation tools and to easily explore their results. All tools apply machine learning to the analysis of evolutionary sequence conservation. Prediction results are shown at a glance on a summary tab, where they are mapped onto the input sequence. Other tabs give further details of the output of individual programs, and all results can be downloaded both in graphic and text format. Below is a very short overview of some predictions you can obtain for the sequence(s) of your own interest. Please check our website to find out more and feel free to email your feedback and requests!

Buchan DW, Minneci F, Nugent TC, Bryson K, Jones DT. Scalable web services for the PSIPRED Protein Analysis Workbench. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W349-57

Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol. 1999 Sep 17;292(2):195-202

MEMSAT-SVM & MEMPACK -

These tools help unveil important features of alpha-helical transmembrane proteins, such as signal peptides, the positions and in/out orientations of membrane-spanning segments, re-entrant helices and pore-lining regions. Additionally, MEMPACK can predict the packing arrangements of transmembrane helices.

Nugent T, Ward S, Jones DT. The MEMPACK alpha-helical transmembrane protein structure prediction server. Bioinformatics. 2011 May 15;27(10):1438-9

DISOPRED3 - http://bioinf.cs.ucl.ac.uk/disopred/

DISOPRED classifies residues as ordered or disordered (lacking stable structure and potentially folding upon binding). It also predicts disordered protein binding sites involved in transient protein-protein interactions.

Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J Mol Biol. 2004 Mar 26;337(3):635-45

FFPRED - http://bioinf.cs.ucl.ac.uk/ffpred/

FFPred assigns GO terms to eukaryotic proteins, which cannot be annotated by similarity but show clear patterns of biological features (e.g. signal peptides, transmembrane helices, secondary structure content, disordered regions, post-translational modification sites etc.) that can be derived from sequence.

Minneci F, Piovesan D, Cozzetto D, Jones DT. FFPred 2.0: improved homology-independent prediction of gene ontology terms for eukaryotic protein sequences. PLoS One. 2013 May 22;8(5):e63754. doi: 10.1371/journal.pone.0063754.

FFSEARCH – http://bioinf.cs.ucl.ac.uk/FFSearch/

Given a set of proteins, FFSearch searches for mouse and human proteins with similar biological features derived from the primary sequence. The output results can be ranked by overall feature similarity or by the likelihood to perform one of the molecular activities or biological processes in a limited, but general enough, subset of GO. Naïve Bayes classification of the database entries is also possible, if users can provide suitable lists of positive and negative examples.

Tutorials

The tutorials have been included in this handout for your reference, however they are available online at:

http://genome3d.eu

Tutorial: FUGUE/VIVACE

Like all methods on Genome3D, the VIVACE pipeline is based on homology modelling, which at its core depends on detecting and using structural similarity to derive information about our systems. What kind of information is most valuable to us will depend on our research, but, structurally, it will invariably come from the homologues we find.

While in some cases we are lucky to find even one, in others we are spoilt for choice. Depending on our goals, the selection might not make much difference, but it may also potentially lead to the disagreements we can observe on Genome3D. In any case, the variety of options itself can often give us valuable insights about the system.

Through its use of the TOCCATA database, the VIVACE pipeline is designed from its conception to be aware of the variety of templates and to try and present it and use it efficiently.

Go to the Genome3D page for uniprot:Q8N427 and click on one of the FUGUE links. It doesn't matter whether SCOP- or CATH-based, they both lead to the same place. That page displays the detailed FUGUE results against the profiles of the TOCCATA database. The first column indicates the profile hit and it shows why the choice of resource made no difference: internally, TOCCATA uses its own SCOP/CATH consensus system to categorise domains, which can be seen in how some profiles are labelled with a nomenclature from SCOP, CATH or both. Unlike Genome3D's own consensus endeavour, the objective was not so much the analysis and comparison of the resources, but to cluster domains from both systems into as few sensible and consistent groups as possible (which is the reason why SCOP was used at the family level instead of superfamily, as you may notice). The length column describes the length of the listed profile. The Z-score column shows the significance of the profile for a given region of the sequence, specified on the last. It is colour-coded for guidance, but basically anything over a value of 6 should be safe to assume as similar and over 4 as likely so.

Click on the name of a profile (a good example would be d.58.6.1-3.30.70.141). It will take you the primary page of a TOCCATA profile. It consists of three sections:

- A JOY formatted alignment of representative sequences: JOY uses typography and formatting to encode structural information into a sequence alignment that can be viewed at a glance to facilitate inspection and comparison of structures, often without needing to visualise 3D structures. The basic original format includes information such as secondary structure, solvent exposure and main chain hydrogen bonding and its key can be seen here, However, the XSuLT extension potentially adds many other features, such as residue depth, inter-residue contacts (not currently displayed), interface and ligand binding residues, as well as data from further such as sequence conservation (entropy) and RMSD from superposition, which highlights areas of spatial variability in a fold.
- A list of related profiles: There are a few ways in which a profile can be related to others. For instance, in this case there is 3.30.70.141. You might notice that the 3.30.70.141 superfamily is also part of the current consensus profile. However, the structures on that profile have not been characterised under SCOP, and TOCCATA, rather than make assumptions about whether they should belong in the consensus, prefers to create an individual profile for them (incidentally, this CATH/SCOP pair happens to be a "Bronze" mapping on Genome3D, since there are some SCOP assignments to the parent superfamily that do not match the CATH one). Another way profiles can

be related, not exemplified in this case, is when one of the elements can be found in any multi-domain patterns present on TOCCATA.

- A list of all domains or chains belonging to this profile: This presents all PDBs in this profile clustered at various identity thresholds and annotated according to their conformational status, in addition to experimental data. Chains with the same cluster number/colour, belong to the same cluster at the given threshold.

The button at the bottom of the page allows you to align any sequence to that particular profile using FUGUE. The resulting page will provide the estimated Z-score for the alignment as well as give the option of customising the template selection (including those from related profiles, if so desired).

Let's go back to the result page for the VIVACE prediction, and go under the "Model and alignments" tab. There we can view an interactive model of each predicted domain along with an XSuLT representation of the alignment used to generate the model. In addition to the features seen on the TOCCATA website, when used on an alignment with sequences or models in addition to structures, XSuLT also includes secondary structure and disorder prediction for the sequence, represented as a coloured line (red for helix, blue for strand, green for disorder) on top of the modelled sequence, which can help to assess the quality of the alignment that is critical to the resulting model.

Tutorial: SUPERFAMILY

This short (~10 min) tutorial will take you through some of the domain assignment and phylogenetic features of the SUPERFAMILY database.

Open your favourite web browser and navigate to the Genome3D website: http://genome3d.eu/

Click on the Search button in the top navigation, then click on Human in the list of filters: http://genome3d.eu/search?page=1&species=human

Here you'll find the first page of all Human proteins in Genome3D, taken from the Uniprot database. Using the search bar at the top of the page, search for "Q3KNS1" and click on the first gene in the results titled PTHD3_HUMAN: http://genome3d.eu/uniprot/id/Q3KNS1/annotations

From here you can see an overview of the predicted domains from the various Genome3D partners. Down the right of the page you can see each of the predicted superfamilies and which structural classification hierarchy they came from (SCOP or CATH). Note the gold, silver & bronze rating, these indicate the degree to which the SCOP and CATH classifications agree.

Look at the first figure, the numbers along the bottom mark the protein sequence position, and the coloured bars show each Genome3D partner's domain predictions.

From the figure we can see three of the partner predictions methods agree on one domain (Multidrug efflux transporter AcrB transmembrane domain), but no obvious consensus for the first domains in the sequence, lets try another protein. Click back to return to the search results and search for "Q3KNS6": http://genome3d.eu/search?q=Q3KNS6

Click on the first result reading Zinc finger protein 829 : http://genome3d.eu/uniprot/id/Q3KNS6/annotations

Here we can see better concensus between the different prediction methods. Look at the SUPERFAMILY domain prediction row; you'll see two domains predicted, a repeat of beta-beta-alpha zinc finger domain and a single Krüppel associated box (KRAB) domain.

To find out a little more about this specific domain arrangement or "architecture" we can jump to the SUPERFAMILY online resource by clicking on the SUPERFAMILY link to the right of the domain predictions: http://supfam.org/SUPERFAMILY/cgi-bin/gene.cgi?genome=up&seqid=Q3KNS6

This is the SUPERFAMILY page listing domains for the Q3KNS6 UniProt? entry for this protein. From here you can see a list of all the assigned domains, the SCOP family and superfamily assignments, and their E-value, along with some links to other information.

Let's see what other species have genes with this identical architecture. Scroll back to the top and click on the third link under the red & blue figure labeled See the phylogenetic distribution for this domain architecture Link: http://supfam.org/SUPERFAMILY/cgi-bin/createtree.cgi?tophl=1;highlight=arc_109640,57667,57667,57667,57667,57667,57667

This tree shows all of the species in SUPERFAMILY contain this archiecture exactly, green branches indicate assignment to the architecture, blue indicates no hit. You can see that our architecture exists in all primates, many mammals, and even the Green Anole lizard Anolis carolinensis. This specific KRAB related architecture appears to have been created in a common ancestor to reptiles, birds and mammals. Of special interest is the level of conservation in primates compared with other mammals, and loss among non-mammals making this perhaps an interesting target for human research.

Let's look at the KRAB Domain in isolation, Press back to navigate back to the list of domains in our domain architecture (with the blue and red demain diagram at the top), scroll down to the 4th domain under the Domain assignment details heading and click on the KRAB Domain link: http://supfam.org/SUPERFAMILY/cgi-bin/scop.cgi?sunid=109640

Here we can see an overview of SCOP classification and functional terms from dcGO[1] specific to the KRAB domain. Notice in the table of Gene Ontology terms you can see many mentions of regulation and negative regulation, indicating this domains known role as a transcriptional repressor. However, for other proteins of novel domain-architecture these terms are predicted using just domain composition and annotation from other hand annotated proteins.

Let's look at the distribution of this domain in our tree of life, click on the Taxonomic Distribution tab at the top of the page, then at the bottom of the following page, click on Plot Tree as SVG: http://supfam.org/SUPERFAMILY/cgi-bin/createtree.cgi?tophl=1;genomes=;highlight=109640

This tree is the same as before, each leaf of the tree is a species, green indicates strong hits against this superfamily for it, blue indicates none. As you can see, we find something unusual, a large subtree of animals where everything is assigned this superfamily, and further down, another large subtree, where just one species has an assignment: Brugia malayi. Let's see if we can figure out what's going on. Click on Brugia malayi in the tree: http://supfam.org/SUPERFAMILY/cgi-bin/info.cgi?genome=r0

On this page you can see a table of assignment statistics, and a larger table showing all superfamily assignments. Lets look for our KRAB domain, search the page (using ctrl+f in your browser) for 'KRAB' and you'll find one hit towards the bottom of the table, click on it: http://supfam.org/SUPERFAMILY/cgi-bin/genome.cgi?sf=109640&listtype=sf&cgi_r0=yes

This page lists each SUPERFAMILY assignment of the KRAB domain to the Brugia malayi species in SUPERFAMILY. You can see a single assignment in the grey outlined box (typically you would see more here) which has an E-value of 9.94e-20, indicating this is unlikely to be a false positive assignment given the sequence.

Press back to navigate back to the genome page: http://supfam.org/SUPERFAMILY/cgi-bin/gen_list.cgi?genome=r0 and click on the Genome information tab at the top of the page: http://supfam.org/SUPERFAMILY/cgi-bin/info.cgi?genome=r0;listtype=sf;subgenome=

This page shows metadata about the Brugia malayi WS218 genome in SUPERFAMILY, including where the genome sequences were sourced, the NCBI taxonomy name, synonyms, and when the genome was added to the database.

Look at the Names and Synonyms section of the table and you will see 'agent of lymphatic filariasis'. A cursory Google confirms that Brugia malayi is a parasitic worm that causes Elephantitis in humans, interesting!

There are two possible reasons for the unusual assignment of the KRAB containing protein outside of Tetrapods:

1: Horizontal gene transfer, perhaps due to the parasitic nature of Brugia malayi

2: Contamination during sequencing

Which one is right? If you think you know, or if you have any other feedback about SUPERFAMILY let us know:

http://supfam.org/SUPERFAMILY/feedback.html

Thanks for participating.

[1] http://supfam.org/SUPERFAMILY/dcGO/

Tutorial: PHYRE2

Here we will study the predicted structure of UniProt Q14654, Gene Name KCN11, the human ATP-sensitive inward rectifier potassium channel 11. In particular the sequence variant Arg 201 to His is associated with neonatal diabetes. There is no X-ray or NMR structure for the human protein, so we require a predicted structure to provide insights into the structural basis for this effect.

1) Analysis using Genome3D

Open your favourite web browser (preferably not Internet Explorer!). Navigate to the Genome3D page for Q14654, either by opening that link or by following the instructions in the next paragraph.

Navigate to the Genome3D website: http://genome3d.eu and on the front page put in the search box the UniProt id Q14654 and run search. You will see "Displaying 1 matching UniProt entry". Click on either the IRK11_HUMAN link or the number 8 beneath "Structural Predictions".

You can see an overview of the predicted domains from the various Genome3D partners. Down the right of the page you can see each of the predicted superfamilies and which structural classification hierarchy they came from (SCOP or CATH). Note the gold, silver & bronze rating, these indicate the degree to which the SCOP and CATH classifications agree.

Look at the top figure, the numbers along the bottom mark the protein sequence position, and the coloured bars show each partner's domain predictions. From the figure we can see from roughly residues 50 to 155, three of the partner prediction methods agree on one domain (Voltage-gated potassium channel) and the CATH and SCOP domain assignments agree; there are no domain assignments from the other partners. For the region 156 – 360, all partners provide predictions, but on closer inspection there are quite different domain assignments. However if you click on b.1.18 on the top right this will take you to SCOP. In the list of families you will find the potassium channel (family number 10).

Return to the Genome3D page. Now look at the section below titled "Predicted 3D Structures". If you click on all the bars and highlight them in orange and then launch a viewer (say PyMOL), you will see the structures superposed. There is good agreement between all the models.

2) Analysis using Phyre2 link from Genome3D

On the right hand side of either the "Predicted Domains" or "Predicted 3D Structures", you will see links to the Genome3D modelling resources. Click on PHYRE2 and this takes you to the details of the predicted Phyre2 model.

The page is divided into four main sections a-d, explained below:

a) Summary. This section displays an image of the highest confidence single model from Phyre2, information on the template (known structure) used to build the model, confidence in homology, coverage of the input query sequence by the model, and an option to view the model using JSMol (this should work on all browsers except Internet Explorer at the moment).

b) Secondary structure and disorder prediction. The sequence has been processed by the programs PSI-Pred and Diso-Pred to predict the locations of alpha helices, beta strands and disordered regions (shown with a question mark). Confidence in the predicted state for each position is shown using a rainbow colour code: red=high confidence, blue=low confidence.

c) Domain analysis. The next table shows what regions of the input sequence have been matched by known structures colour coded by confidence in the match. This enables you to see the approximate domain structure of your protein.

d) Detailed template information. This is the main table of results showing a ranked list of matches to known structure (the template), information about the template, the region aligned and an image of the model produced. Clicking on the protein picture will download a PDB formatted file of the model of your protein based on the template shown. For even further detail, click on the “Alignment” button in a particular row.

As Genome3D is a database of pre-computed models, if the Phyre2 information is valuable to you, the best procedure is to rerun Phyre2 from its web page to obtain the very latest predictions with the most recent analysis features.

For the purposes of this workshop, we have recently re-run UniProt Q14654 through a newer version of Phyre2. The results of this analysis can be found by scrolling to the top of the results page and clicking on a link entitled: “Click here for UPDATED RESULTS for Genome3D Workshop”. Click that link now.

In general, if you want to resubmit a sequence to Phyre2, just visit the main submission page. This can be found here: http://www.sbg.bio.ic.ac.uk/phyre2/ or just by searching for Phyre2 in Google.

3) Phyre Investigator

Once you have clicked on the “UPDATED RESULTS” link in Step 2, you will be taken to a newer version of the analysis. The layout is largely similar to that seen in Step 2 but with optional show/hide links to avoid screen clutter. If you go to the “Detailed Template information” section there is, in each row, a new link in the rightmost column saying “View Investigator results”. (Normally a button would be present here for you to choose whether to run Phyre Investigator, but we have already run all these analyses for you.)

Go to the first entry in the table of results (template c3syaA_). Click on the “view Investigator results” link. This will take you to a page showing the results of running a large number of different analysis programs on the top-ranking model from Phyre2.

From left to right, the page displays the model in an interactive JSMol window, buttons providing a choice of analyses, and two bar graphs displaying preferred amino acids at each position in the sequence and mutations likely to have a phenotypic effect as predicted by SuSPect.

At the bottom of the page is a sequence and secondary structure view. As you hover your mouse over a region of the sequence, it will highlight where that residue is in the model. Scroll the sequence pane to the right to find residue R201. If you hover over that residue in the sequence view you can see the mutations graph above shows that almost all changes to this residue (with the exception of an R->A mutation) are strongly predicted (tall red bars) to have a phenotypic effect.

If you click on this residue it will spacefill that position in the JSmol window. Go to the “Analyses” panel, click on “Function” and click on “Mutational sensitivity”. This colours the entire structure by the average confidence from SuSPect of a mutation having a phenotypic effect. As you can see, the highlighted R201 is coloured orange indicating high sensitivity to mutation.

4) The SuSPect web server

SuSPect is our predictor of the phenotypic effect of amino acid variants including nsSNPs. It uses a machine learning approach (an SVM) based on sequence and protein-protein interaction network features to make the prediction. Within SuSPect there are either pdb or Phyre2-predicted structures stored.

To explore the use of SuSPect click on the link:

http://www.sbg.bio.ic.ac.uk/~suspect/

You can enter Q14654 R201H and run to see the effect of this mutation

You will see a score of 87 suggesting that this sequence change is disease associated. Click on this score and you see the features used in SuSPect to make this prediction.

References

Kelley, L.A. & M.J. Sternberg, Protein structure prediction on the Web: a case study using the Phyre server. Nat Protoc, 2009. 4: p. 363-71.

Yates, C.M., I. Filippis, L.A. Kelley & M.J. Sternberg, SuSPect: Enhanced Prediction of Single Amino Acid Variant (SAV) Phenotype Using Network Features. J Mol Biol, 2014. 426: p. 2692-701.

Note: We recommend you open this page's links in new tabs (hold down the Ctrl key whilst clicking the link or click the link with the right mouse button and select Open Link in New Tab) so that you don't navigate away from this page.

To help keep the tutorial short, we have provided short-cut links to complement the instructions. We recommend using Firefox or Chrome rather than Internet Explorer.

A tour through CATH and a story of survival

You are interested in a beta lactamase protein that you know confers drug resistance in a pathogenic bacterium and you want to find out more about the nature of the active site and why this pathogen is resistant to penicillin. Search Genome3D (www.genome3d.eu) with the UniProt identifier P00811. You will see that the protein belongs to CATH superfamily 3.40.710.10 of beta-lactamases and DD-peptidases. The majority of Genome3D structure prediction methods assign the protein to equivalent superfamilies in SCOP and CATH.

Click on the Gene3D link and then on Click Here to get more information from the CATH-Gene3D resource (or click here if you get lost). Click on the Drugs entry of the menu on the left of the page to see information available on the drugs found to bind to the active site of this protein. There is one approved synthetic drug and many other putative drugs for which there is experimental evidence. Go back to the top of the page and take note of the fact that P00811 is in Functional Family 22208 of beta-lactamases and that this is associated with the EC code 3.5.2.6. We will investigate Functional Family 22208 later in the tutorial.

Now let's look at the CATH superfamily (3.40.710.10). Click on CATH superfamily (or here). This page shows summary information for superfamily 3.40.710.10. Mouse-over the Species Diversity box to see the species distribution of the superfamily. You will see that this superfamily is largely confined to bacterial species. Now mouse-over the EC Diversity box to see the functions associated with different relatives in the superfamily. Apart from beta-lactamases, what is the largest functional group?

What about the evolution of the proteins?

To find out when these different functions emerged in evolution, click on the FunTree box (under Enzyme Function). You can see that FunTree provides a lot of phylogenetic and functional information for the superfamily. This is organised according to sets of similar structural groupings (SSG). Structures in this superfamily are quite similar (ie superpose within 9Å RMSD) and there is only one SSG in the superfamily. Click on the link for this SSG (which is an image of a structure to the right of SSG on the left of the screen; click here if you get lost) and then select the Reduced Phylogenetic Tree Showing EC Changes. This shows the earliest point in evolution at which a particular function can be detected within the domain superfamily. The beta lactamases (with EC code 3.5.2.6, as seen earlier) are clearly ancient and evolved at about the same time as the largest group of relatives in the superfamily, the DD-peptidases (EC code 3.4.16.4).

The DD-peptidases (or "D-Ala-D-Ala carboxypeptidases" in the EC diversity box, earlier) are bacterial enzymes that are also known as "penicillin binding proteins". They cross-link peptidoglycan chains to form a strong mesh-like structure in bacterial cell walls. The figure to the right describes their activity in Gram-positive bacteria.

These enzymes are serine peptidases i.e. during catalysis an acyl-enzyme intermediate is formed involving a serine residue. The use of D-amino acids by bacteria contributes to the resilience of their cell walls since proteins only contain L-amino acids and most peptidases are L-stereospecific.

About 2 billion years ago fungi evolved the ability to synthesize beta-lactam antibiotics which bind irreversibly in the active site of DD-peptidases and thus inhibit their activity. Because bacterial cell walls are constantly being broken down and remodelled by the bacteria themselves, an inability to regenerate cross-linkages rapidly results in loss of integrity of the cell wall and bacterial death. An example of a current day organism is the penicillin mould that produces penicillin and in 1928 this was the first antibiotic to be discovered by Man.

In response to this assault it is generally accepted that some bacterial DD-peptidases evolved into beta-lactamases that can break open the beta-lactam ring in an antibiotic and thus inactivate it.

How has resistance evolved? Can we see changes in the active site?

Let's look more closely at these two different relatives. Go back to the CATH summary page. We can use the functional family (FunFam) classifications in CATH, which are listed in tree format the bottom left. FunFams are clusters of relatives that are very likely to have similar functions. Click on Alignments in the middle of the SUPERFAMILY LINKS menu at the top left. Find FunFam 22208 (Beta-lactamase [FF: 22208] in the table and open the link in a new browser tab. This FunFam contains our original beta-lactamase, P00811.

Select the Alignment tab to view the multiple sequence alignment of this functional family. Highly conserved positions have been identified by the Scorecons algorithm ("Scoring residue conservation" Valdar WSJ (2002), Proteins: Structure, Function, and Genetics. 43(2): 227-241) and are shown highlighted in green. Can you spot the SXXK motif associated with the catalytic dyad used by this family? It should be around alignment positions 62-65.

Now switch back to the browser tab containing the table of FunFams, open FunFam 22165 (D-alanyl-D-alanine carboxypeptidase [FF: 22165]) in a new browser tab and click on the page's Alignment tab. FunFam 22165 contains DD-peptidase relatives. You will see a similar motif as both enzymes possess the SXXK motif catalytic dyad (additional catalytic residues are often suggested but these are also shared by both classes of enzyme). This should be around alignment positions 45-48. So the difference in their catalytic activity is likely to be due to substrate binding residues that perhaps subtly alter the position of the substrate, alter the energetic stability of binding, stabilization of the transition state intermediate and/or subsequent hydrolysis of the acyl-enzyme intermediate.

Can you see any highly conserved residues which lie close to the SXXK motifs but which are different in the beta-lactamase and DD-peptidase alignments? For example, the following positions involve quite significant changes: 67 in 22208 versus 50 in 22165 and 70 in 22208 versus 53 in 22165.

How close are these residue mutations?

CATH provides two structural comparison resources. The CATHEDRAL web-server allows users to submit a structure to be scanned against a library of representative CATH domain structures. The SSAP web-server allows a quicker structural alignment and superposition of two structures.

To save time in this tutorial, we have performed these searches for you. Click here to see the search of the 1my8A00 beta-lactamase structure against CATH. The top hit is for another beta-lactamase within the same superfamily (which you can see by clicking on the domain ID and then on the domain's FunFam entry). Now let's compare the structure of the beta-lactamase with a structure from a DD-peptidase family. Click on this PyMOL superposition (which was pre-prepared using the SSAP web-server mentioned above) of protein structure 1my8A00 for the beta lactamase P00811 against the structure 3itbD01 for the DD-peptidase P08506. How similar are the structures? The residues involved in the SXXK motif in the active site are highlighted in green. How similar are the active site regions of these functionally diverse domains?

Here are some residue positions that are highly conserved within each FunFam but different between the two FunFams:

```
22208 (1my8A00) : ...SXXK.F..V...

                          ↑  ↑

22165 (3itbD01) : ...SXXK.M..Y...
```

You can verify that each of these is highly conserved in the two FunFam alignment browser tabs you now have open. Let's select these residues in PyMOL: look at the two sequences at the top of the PyMOL window, find the SXXK motif (highlighted in green) and then, for both sequences, select the residues two and five residues after the K. (Note: the two sequences aren't aligned in PyMOL so these residues are in different positions.) Type show sticks, sele and then look closely at the side chains of these residues. Would you say that these mutated residues are likely to have an impact on the binding of the peptide substrate?

Some thoughts to take away…

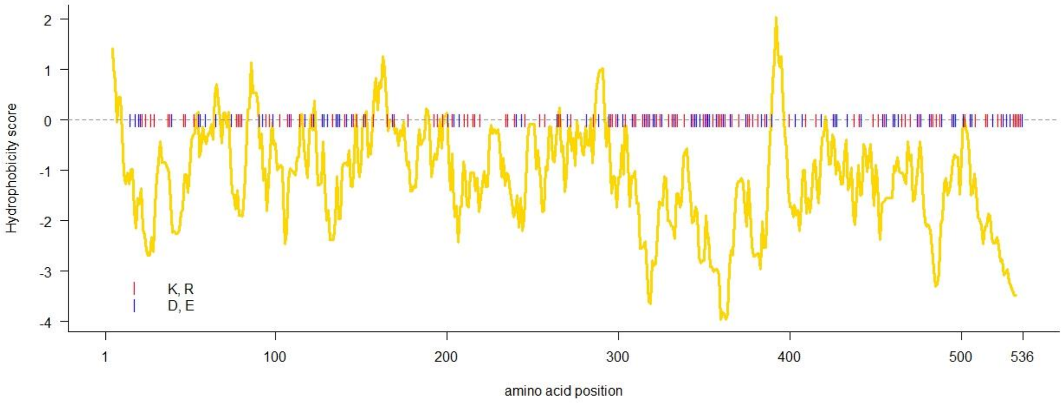
Novel beta-lactamases rapidly evolve and confer resistance against new Man-made derivatives of beta-lactam antibiotics (typically within two years) as is the case for all other existing classes of medicinal antibiotics, resulting in what may be an imminent crisis for modern medicine. People may start dying from scratches again as they often did in the pre-antibiotic days of previous centuries. Pharmaceutical companies are reluctant to invest in the development of new classes of antibiotics since they would need to be kept in reserve for emergencies and thus be infrequently used and result in little revenue.

Prediction of intrinsically disordered regions and protein binding sites within them

Sometimes you will find no reliable three dimensional model of the protein structure you are interested in. This may be due to a number of different reasons, such as the lack of suitable templates or the limits of fold recognition methods to identify them. In the case of intrinsically disordered proteins or protein regions, the amino acid chain haa a relatively flat energy landscape, and so it samples a broad set of conformations that cannot be reasonably approximated by a single structure. Let's have a look at the Genome3D entry for the human SNW domain-containing protein (also called nuclear protein SkiP) - please open this link in a new tab or page: Q13573

The 'Predicted 3D structure' section in the 'Annotation' tab shows that only one prediction by the VIVACE pipeline is available, but it covers only 80 residues out of 536 as a long helix. Regardless of the accuracy of such prediction, clearly it is not that useful to figure out the biological role of this protein.

In fact, SkiP appears to be largely disordered in vivo and the figure below highlights some distinguishing features. The golden line shows the average hydrophobicity score calculated over a window of 9 residues; positive values indicate more hydrophobic regions (that would promote folding), while negative values correspond to more hydrophilic segments. Charged amino acids are also highlighted with red (positively charged) and blue (negatively charged) vertical bars. The high proportion of hydrophilic and charged residues is typical of intrinsically disordered proteins and regions.



Let's use DISOPRED to analyze the amino acid sequence of SKiP, obtain residue-level disorder predictions and detect protein binding sites within them. Click on the 'PSIPRED' button under the 'Associated tools' heading and you will be redirected to submission form of the PSIPRED webserver. Please make sure you tick the 'DISOPRED3 and DISOPRED2' box for this tutorial. You are welcome to run the available tools against other proteins in the future, but we kindly ask you not to run other analyses during this tutorial due to the limited time available. Press 'Predict' to run the job.

The results summary page recaps the prediction output and maps it onto the input sequence as per the key. At the top of the page you can usually see the short identifier you provide for the job and the unique private ID assigned by the PSIPRED server. Below this, separate tabs allow you to view the specific outputs for each tool that was run. For the purpose of this tutorial, all analyses have been run in advance and the different tabs show the pre-calculated results; the identifier at the top of the page is therefore set to 'test_Q13573'.

The large majority of amino acids is predicted to be disordered (and highlighted in red boxes), and some are also likely to bind other proteins (shown in green boxes). The ‘DISOPRED‘ tab gives a graphical representation of more detailed information about disorder predictions. The disorder profile plot shows the DISOPRED3 disorder confidence levels against the sequence positions as a solid blue line. The grey dashed horizontal line marks the threshold above which amino acids are regarded as disordered. For disordered residues, the orange line shows the confidence of disordered residues being involved in protein-protein interactions. Disordered amino acids are predicted to form protein binding sites when the confidence scores are larger than 0.5. The 'Downloads' tab allows you to save locally the results of the analysis both in graphical and text format.

How reliable are these data? Using NMR spectroscopy, a recent study showed that positions 1-172 are disordered in isolation and that the segment spanning positions 59-79 folds upon binding the protein PPLI (Wang X et al. "A large intrinsically disordered region in SKIP and its disorder-order transition induced by PPIL1 binding revealed by NMR." J Biol Chem. 2009). DISOPRED correctly classifies approximately 65% of the 172 N-terminal disordered residues, and this approximately mirrors the accuracy levels achieved during the independent CASP benchmarking experiments. Given the lack of experimental data, we cannot make defintive statements about the quality of the other predicted disordered regions, but these appear to be consistent with common assumptions and with consensus data in external resources such as MobiDB and D2P2. The disordered protein binding site from position 59 to 79 is predicted with 38% precision.

Prediction of protein function from sequence (FFPred), and helical packing arrangement for transmembrane proteins (MEMPACK)

In order to try some other tools among those provided by the main PSIPRED server, let's now consider a very different protein.

Human Rhodopsin is one of the proteins responsible for the perception of light in our species. It is a transmembrane protein, belonging to the G-protein-coupled receptor (GPCR) family.

As usual, let's start by visiting the Genome3D page for Rhodopsin by opening this link in a new tab or page: P08100

Click on the 'Annotations' tab. The page summarizes the structural information in the way we are already familiar with. Once again, we can retrieve the PSIPRED server home page by clicking on the 'PSIPRED' button under 'Associated tools' at the bottom of the page.

Notice how the PSIPRED server page has been already filled with the protein's amino acidic sequence. The tool is ready to run. As before, we will visit a cached result for this protein during this Workshop. In real usage, you need to remember to provide a short identifier for your PSIPRED jobs ('Short identifier for submission'); today, the identifier will be automatically modified to 'test_P08100'. Please remember that you are kindly requested not to alter the submission sequence for the purpose of this tutorial - however, please feel free to use these tools as you wish in the future.

The PSIPRED suite contains a tool for prediction of protein function directly from amino acidic sequence, using limited or possibly no homology information. This tool is called FFPred, and its latest version (v2.0) can be used if you click on the corresponding checkbox under 'Choose Prediction Methods'. Moreover, the suite includes a tool for predicting the transmembrane helical packing arrangement (MEMPACK): please select this checkbox as well. You may also want to de-select the PSIPRED checkbox, that is usually ticked by default.

After clicking 'Predict' you are redirected to the cached result for this protein (actually running the job would be time consuming). On the results page, the tabs we are interested in are those named after the tools we just mentioned.

The 'FFPred' tab shows the output of FFPred 2.0 for human Rhodopsin. The top section of the output includes two tables, for the two Gene Ontology (GO) domains of 'Biological Process' and 'Molecular Function'. The tables list a series of "GO terms" belonging to those GO domains, that have been predicted to be annotated to Rhodopsin; GO terms are the standard way to annotate functional characterisation to proteins. Each line in the tables contains the GO term and its description, followed by the posterior probability of the prediction being correct and, finally, an indication of the overall reliability, high (H) or low (L), of that particular GO term.

In order to understand the output, we need to know a bit more in detail how FFPred achieves its predictions. The input protein sequence (Rhodopsin in our case) is first analysed by FFPred, which runs a series of prediction tools and extracts biologically relevant "features" of the protein - for instance, the number of alpha helices (if any), the average hydrophobicity of the protein and many more. We'll see this in more detail in the next paragraph. Then, for each different GO term in its vocabulary, FFPred runs a Support Vector Machine (SVM) on Rhodopsin's set of features, and by doing so it compares Rhodopsin with sets of proteins for which the functional characterisation for that particular GO term is known. This allows FFPred to give back the probability (indicated in the tables) that Rhodopsin actually is annotated with each GO term - the higher the probability, the "safer" the prediction.
Finally, GO terms that are in general very harder to predict, and therefore always included only as speculations on possible functional characterisation, are included as low (L) reliability predictions, on a red background, while all other GO terms are considered highly reliable (H), and are always shown at top of table, regardless of the predicted probability. This simply means that "red" GO terms are *always* to be considered less reliably predicted than the others, even though they may be predicted with a high probability for this particular protein, as happens for "cellular protein modification process" in this case.

In our case, looking at highly reliable GO terms only, we can see how Rhodopsin is correctly predicted to be annotated with GO terms like "G-protein coupled receptor signaling pathway", "detection of stimulus" (Biological Process), "G-protein coupled receptor activity", "signal transducer activity" (Molecular Function) with very high probability values. This should not be surprising, as Rhodopsin itself is a well-known example of such properties and this has therefore been recognised by the corresponding SVMs. Other examples of predicted GO term annotations however have lower probabilities, and can be interpreted as suggestions of further functional characterisation that may be tested in experimental work.

The bottom section of the 'FFPred' tab includes indications of which "features" of Rhodopsin's sequence have been used to obtain the predictions. Features range from structural features (remember, however, that no structural data is used: these are computationally predicted features), disorder, post-translational modifications, PEST regions, amino acid composition, physico-chemical properties of the protein. In particular, some of these are actually predicted using some of our other tools - PSIPRED is used to predict secondary structure, DISOPRED is used for disorder and MEMSAT-SVM is used to predict transmembrane helices and their topology.

Lastly, let's focus on the transmembrane helices that are predicted for Rhodopsin. You can see the 7 predicted helices both in the cartoon on this tab, as mentioned in the previous paragraph, or directly clicking the results tab for MEMSAT-SVM, which is the tool that was used to obtain such prediction. It is well known that Rhodopsin, as a GPCR, exhibits 7 transmembrane helices - however, can we say any more about the arrangement of these helices in the lipid bilayer?

Clicking on the 'MEMPACK' results tab shows the prediction for the Rhodopsin helical packing arrangement made by this other tool within the PSIPRED server suite. The seven predicted helices are depicted in a diagram in the most likely predicted conformation. Lines connecting residues on the different helices represent interactions that are thought to make such conformation stable and the most favoured one.

This diagram can be downloaded directly by clicking on it; many more useful analysis files for Rhodopsin, including those containing FFPred output details, can be downloaded by visiting the 'Dowloads' tab of these results pages.

Resource Presentations