Annotating Genomes with Structures and Functions

BBSRC funding from 2011, Established July 2012
SAB: Prof Geoff Barton – Dundee University,
Prof Chas Bountra - Structural Genomics Consortium,
Prof Torsten Schwede - Swiss Institute of Bioinformatics

http://www.genome3d.eu

Groups involved in Genome3D

Tom Blundell – Cambridge University
Julian Gough – Bristol University
David Jones – UCL
Alexey Murzin – LMB (Cambridge)
Christine Orengo – UCL
Michael Sternberg – Imperial, London

Resources

SCOP (Murzin) ~3000 domain structure families
CATH (Orengo) Predicted domain annotations for >30 million sequences in UniProt,
SUPERFAMILY (Gough) ~70% of domains in completed genomes
Gene3D (Orengo) Predicted domain annotations and 3D models for selected organisms
FUGUE (Blundell)
PdomTHREADER (Jones)
PHYRE (Sternberg)

Domain Structure Classification

<table>
<thead>
<tr>
<th>SCOP</th>
<th>CATH</th>
<th>Consensus</th>
</tr>
</thead>
</table>

Domain Structure Annotation

- SUPERFAMILY
- Gene3D
- PHyre
- pDomTHREADER
- FUGUE

Associated functional information from > 10 public sources:
- KEGG terms
- GO terms
- Other...

ELIXIR

ELIXIR unites Europe's leading life science organisations in safeguarding the biological data generated every day in publicly funded research. Learn more at www.elixir-europe.org

Predicted Domain Annotations (Per UniProt Sequence)

- # groups
- # groups
- # groups
- # groups
- # groups
- # groups
- # groups
- # groups
ELIXIR-UK

- UK node’s initial focus will be exclusively on training
- The UK node will develop training infrastructure and focus on:
  - training needs analysis and trainer workshops
  - e-support service platform (TeSS)

Genome3D Applications

- SCOP
- CATH
- Consensus
Organisation of Information

CREDO
database of protein-ligand interactions

- represents contacts as structural interaction fingerprints,
- sequence-to-structure mapping
- molecular shape descriptors with Ultrafast Shape Recognition (USR),
- fragmentation of ligands in PDB,
- identification of approved drugs.
- completely scriptable through application programming interface.

FUGUE

- Sequence-structure homology recognition program

- Defining characteristics:
  - Use of Environment-Specific Substitution Tables (ESSTs) in structural profiles
  - Automatic alignment algorithm selection with structure-dependent gap penalties

Structural Environments

- Residues exist in variety of environments in protein structures, this affects their conservation in evolution.
- Examples of environments:
  - secondary structure,
  - solvent exposure,
  - hydrogen bonding of main or side chain,
  - atypical dihedral angles.
- BLOSUM-like substitution tables can be derived for each combination of environments (currently 64), improving the detection of remote homology and alignment quality.

TOCCATA

- Substitutes original HOMSTRAD database as source of profiles for FUGUE.
- Constructed from a consensus of SCOP 1.75(b) families and CATH 3.5 superfamilies, including multi-domain patterns (not used on G3D).
- Goal was to group domains/structures in minimal number of categories, not analysis.
- Each structure annotated according to conformation (ligand binding, oligomeric state).

http://structure.bioc.cam.ac.uk/toccata/

TOCCATA in numbers

- 57,880 PDBids
- 135,894 PDB chains
- 228,014 domains
  - 114,647 with non-trivial ligands
  - 148,605 as part of complexes
- 8151 profiles
  - 6238 single domain families (2263 consensus)
  - 1519 multi-domain profiles
  - 394 repeated domain profiles

VIVACE Pipeline

Genome3D Stats

<table>
<thead>
<tr>
<th>Genome</th>
<th>FUGUE (SCOP)</th>
<th>FUGUE (CATH)</th>
<th>VIVACE (models)</th>
</tr>
</thead>
<tbody>
<tr>
<td>E. coli</td>
<td>3,709</td>
<td>3,642</td>
<td>N/A</td>
</tr>
<tr>
<td>S. cerevisae [baker’s yeast]</td>
<td>5,499</td>
<td>5,430</td>
<td>N/A</td>
</tr>
<tr>
<td>H. sapiens [human]</td>
<td>15,620</td>
<td>14,967</td>
<td>15,133</td>
</tr>
</tbody>
</table>

Joy / XSSuLT

- Encodes structural environment information [e.g. that used by FUGUE]
- XSSuLT expands the original JOY to include other features
  - inter-residue contacts, residue depth, interface & ligand binding residues,
  - predictions & custom per-residue annotations, among others.

Genome3D Stats

<table>
<thead>
<tr>
<th>Genome</th>
<th>FUGUE (SCOP)</th>
<th>FUGUE (CATH)</th>
<th>VIVACE (models)</th>
</tr>
</thead>
<tbody>
<tr>
<td>E. coli</td>
<td>3,709</td>
<td>3,642</td>
<td>N/A</td>
</tr>
<tr>
<td>S. cerevisae [baker’s yeast]</td>
<td>5,499</td>
<td>5,430</td>
<td>N/A</td>
</tr>
<tr>
<td>H. sapiens [human]</td>
<td>15,620</td>
<td>14,967</td>
<td>15,133</td>
</tr>
</tbody>
</table>
Human Genomes & Mutations

Can chemistry, structure and genomics information help identify mutations that cause disease?

Genome Sequences

Mendelian Inherited Diseases BHD Syndrome

Single Gene Mutations & Disease

Early Onset Breast Cancer BRCA2

Polygenic Disorders & Complex Diseases Cancer Somatic Mutations

Which are the "drivers", and which are the "passengers"?

SDM: Stability score calculation

Unfolded state represented by substitutions occurring outside of regular secondary structure, solvent exposed and non-hydrogen bonded


mCSM

Predicting the effect of mutations in proteins using graph-based signatures

http://bleoberis.bioc.cam.ac.uk/mcsm/

Douglas E. V. Pires

mCSM: predicting the effects of mutations in proteins using graph-based signatures. Bioinformatics 30(3):335-342

Genome3D Resources at Biochemistry, Cambridge

Tom L Blundell
Bernardo Ochoa M
James Smith
Department of Biochemistry
University of Cambridge
Superfamily
HMM library and genome assignments server

http://supfam.org

A Hidden Markov model profile

PDB files
SCOP
domains, classification
SUPERFAMILY
profile HMMs
SUPERFAMILY
genome annotation

Structural assignments to genomes
Superfamily
HMM library and genome assignments server

- ~5,000 genomes
- ~125 million sequences
- GO annotation of domains and sequences
- Phylogenetic reconstruction
- Comparative genomics/enrichment tools
The Phyre2 Web server for protein structure prediction incorporating the Su SPect amino-acid variant phenotype predictor

Lawrence Kelley
Michael Sternberg
Chris Yates
Stefans Mezulis
Imperial College London

How does Phyre2 work?

• “Normal” Mode
• “Intensive” Mode
• Advanced functions

Phyre2 (normal mode)

ARDLVPIMYC
HMM
HMM-HMM Matching (HHsearch, Soeding)
3D Model
ARDL--VIPM--IY
AFDL--CDLVIP--CMAY
Sequence of known structure

Example results

Top model info
Secondary structure/disorder
Domain analysis
Detailed template information

How does Phyre2 work?

• “Normal” Mode
• “Intensive” Mode
• Advanced functions

Shortcomings of ‘normal’ Mode

• Because of local alignment or novel domain combinations, domains often modelled separately
• Regions with no detectable homology to known structure unmodelled
• Does not use multiple templates which, when combined could result in better coverage

Thus need a system to fold a protein without templates and combine templates when we have them
**Poing - simplified folding model**

- **Structure simplification**
- **Backbone C-alpha**
- **Protein backbone**
- **Large hydrophobic sidechain**
- **Small hydrophilic sidechain**

Based on Levitt and Warshel

**Phyre + Poing**

- ARNDSL DVCS......
- PSI-Blast
- Hidden Markov Model DB of KNOWN STRUCTURES
- Extract pairwise distance constraints

**How does Phyre2 work?**

- **“Normal” Mode**
- **“Intensive” Mode**
- **Advanced functions**

**Advanced functions**

- **PhyreAlarm** – automatically re-run tricky sequences every week
- **BackPhyre** – compare a structure to up to 30 genomes
- **One-To-One Threading** – use specific PDB for model building

**PhyreAlarm**

- User sequence: SVYDAAAQLTADVKKD......
- Newly solved PDB Structures added WEEKLY
- Email results New 3D model
- Perform full Phyre modelling
- Newly added Structure HMMs
- HMM-HMM matching
- Confident hit?
- Yes
- No
- Try again next week

**BackPhyre**

- User structure: SVYDAAAQLTADVKKDLRDSW KVIQDSTKLKDVCMQITFAC LOGTSGTKVAKLHGDNLK KLRHSTLVMAQNTFIELDD

Hidden Markov Model DB of Genomes

Rank No. Similar Score
1 GI......
2 GI......
3 GI......
- - -

Ranked list of genome hits

HMM-HMM matching
One to one threading

User structure

SUDDAAAAQQEADVNLCS…….

HMM of User structure

HMM-HMM matching

User sequence

KLIGITSHSLQNLPSOGLCVS…….

HMM of user sequence

Final model

New: Phyre Investigator

- Model quality assessment
- Location of functional sites
- Effect of mutations on structure and function
- Protein-protein Interface(s)

Phyre Investigator

- Clashes
- Rotamer outliers
- Ramachandran outliers
- ProQ2 model quality assessment
- Alignment confidence (HHsearch)
- Conservation/evolutionary trace (Jenson-Shannon divergence –far faster and just as accurate as ET)
- Catalytic Site Atlas
- Disorder
- Pocket detection (Fpocket)
- Protein interface residues (PI-Site, ProtinDB)
- Conserved Domain Database ‘conserved features’ for NCBI-curated domains

Phyre Investigator

Effect of Mutations?

- Will a SNP effect my protein’s function?
- New method: SuSPect
- Recently developed by Chris Yates in our lab
- Integrated into Phyre Investigator
- Also standalone server

Phyre Investigator

Phyre Investigator

Phyre Investigator

Phyre Investigator
SuSpect – Phenotypic effect of amino acid variants

Sequence conservation
• PSSM
• Pfam domain
• Jensen-Shannon entropy

Structural features
• Predicted solvent accessibility

Network features
• Protein-protein interaction (PP as domain centrality)

SuSpect – Results on non-training data (VariBench)

Specificity = $\frac{TP}{TP + TN}$

Sensitivity = $\frac{TP}{TP + FP}$

Benchmark consists of 20k SNPs (15k Neutral, 5k pathogenic)

Neonatal diabetes

• Arg 201 His in ATP-sensitive inward rectifier potassium channel 11 (Kir6.6)
• SuSpect gives score of 87/100 – high probability of disease associated

Phyre2 yields model which suggest structural basis for disease

Most variants predict to be disease associated

Arg 201 forms H-bond with main chain O

His in variant could not form similar interaction
Orengo and Thornton (1993)

HUP superfamily

~280,000 domain structures in CATH

2735 domain structure superfamilies

CATHEDRAL

CATHs Existing Domain Recognition Algorithm

Redfern et al. PLOS Comp. Biol. (2009)

• Rapid graph theory secondary structure filter
• Double dynamic programming for accurate residue alignment

Fold Recognition Performance of CATHEDRAL


CATHEDRAL server

Domain superfamily

• shared topological core 3D-motif
• sequence or functional similarity

Domain structure annotations
CATH-Gene3D: Domain structure annotations

protein sequences from genomes and Uniprot

scan against HMMs for CATH Pfam

> 26 million domain sequences assigned to CATH superfamilies

< 100 superfamilies (<5%) account for 70% of domain sequences in CATH-Gene3D

CATH-Gene3D functional families - FunFams

Large, diverse superfamily

Arg-tRNA synthetase

Pantoate - β-alanine ligase

Asn synthetase B

> 40,000 sequences, > 250 GO molecular function terms

FunFHMMer: Uses Specificity-Determining Positions (SDP)

Recall

Precision

MFO

mode1

Performance in Function Prediction

CAFA international assessment, July 2014

Recall - Precision

Performance in Function Prediction

CAFA international assessment, July 2014

Recall - Precision
FunFams are structurally and functionally coherent

3D models built for human using MODELLER algorithm (Sali Group)

FunFams are structurally and functionally coherent

functional sites predicted in all FunFams using the scorecons algorithm (Valdar and Thornton)

significant enrichment of Catalytic Site Atlas (CSA) residues in conserved residues – 2.2e-16

Dessally et al. BBA (2013)
Structure and sequence alignments for enzyme families → Phylogenetic trees

Annotate with functional information eg small molecule data - substrates, mechanisms etc
A Brief Overview of the PSIPRED Workbench

David Jones
UCL Depts. of Computer Science and Structural and Molecular Biology

The PSIPRED Workbench

- First available to public in 1998
- Originally offered secondary structure prediction (PSIPRED), fold recognition (GenTHREADER) and transmembrane topology prediction (MEMSAT)
- Now covers a range of applications, including protein function prediction, disorder prediction
- Most tools available for download or can be run on the cloud via HADOOP

Scalable web services for the PSIPRED Protein Analysis Workbench
DWA Buchan, F Minneci, TCO Nugent, K Bryson, DT Jones (2013)
Nucleic acids research 41 (W1), W349-W357

pDomTHREADER

- Highly sensitive domain-based fold recognition method
- Outperforms HHPred and PRC on domain superfamily recognition
- Adaptation of pGenTHREADER profile-profile comparison algorithm for local matches
- Part of PSIPRED Protein Sequence Analysis Workbench

ProTera:
- Modular architecture based on structural alignments, secondary structure, potentials of mean force
- Updated weekly
- Tightly integrated with CATH domain releases

pDomTHREADER Domain Superfamily Recognition Performance

Functional implications of protein disorder

At the cellular level, IDPs and IDRs have been linked to the organization and re-wiring of protein-protein interaction networks, and to increased proteome diversity through alternative splicing across tissues and organisms

At the protein level, IDR length and position correlate with their biological roles. Based on this observation, FFPRED assigns GO terms to gene products with limited or no sequence similarity at all to experimentally characterized proteins

Individual IDRs act as flexible linkers or contain binding sites (for proteins, nucleotides, lipids or metal ions) that usually fold upon binding
DISOPRED3 extends its predecessor with the aim of improving prediction accuracy, especially for long IDRs.

Additional modules include:
- a Neural Network trained on a much larger PDB + Disprot dataset;
- a profile-based nearest neighbour method against PDB + Disprot;
- a Neural Network to integrate the results of the component methods.

DISOPRED3 vs DISOPRED2
- Based on CASP data, we found DISOPRED3 to be more specific than DISOPRED2, and more sensitive to long IDRs and far from the N- and C-terminus.
- DISOPRED3 was ranked at the top or near the top across a range of test conditions and evaluation measures by the CASP9 and CASP10 assessment teams.

CASP10 assessment results
- The best methods attain 70% accuracy.
- Internal IDRs are harder to detect than terminal ones.
- IDRs with 40 or more residues remain challenging, but larger test sets are needed.

Benchmarking
- Test set: 9 protein chains from DisProt with less than 30% identity to DISOPRED3 training data, containing 14 IDRs of length between 5 and 50 residues and annotated as protein binding.
- Evaluation measures: precision and recall, as most negatives are expected to be structured residues.
- Naïve predictions: random labelling of input amino acids as either disordered protein binding or not with equal probability.

<table>
<thead>
<tr>
<th>Method</th>
<th>TP</th>
<th>FP</th>
<th>FN</th>
<th>Recall</th>
<th>Precision</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>DISOPRED3</td>
<td>78</td>
<td>324</td>
<td>191</td>
<td>0.259</td>
<td>0.194</td>
<td>0.232</td>
</tr>
<tr>
<td>Naïve</td>
<td>134</td>
<td>1704</td>
<td>135</td>
<td>0.498</td>
<td>0.073</td>
<td>0.127</td>
</tr>
<tr>
<td>DISOPRED3 no PB</td>
<td>81</td>
<td>1681</td>
<td>188</td>
<td>0.301</td>
<td>0.046</td>
<td>0.080</td>
</tr>
<tr>
<td>SPSSMpred</td>
<td>18</td>
<td>306</td>
<td>251</td>
<td>0.067</td>
<td>0.056</td>
<td>0.061</td>
</tr>
<tr>
<td>MoRFpred</td>
<td>16</td>
<td>291</td>
<td>253</td>
<td>0.059</td>
<td>0.052</td>
<td>0.056</td>
</tr>
<tr>
<td>ANCHOR</td>
<td>21</td>
<td>1481</td>
<td>248</td>
<td>0.078</td>
<td>0.014</td>
<td>0.024</td>
</tr>
</tbody>
</table>

CASI10 assessment results
- The best methods attain 70% accuracy.
- Internal IDRs are harder to detect than terminal ones.
- IDRs with 40 or more residues remain challenging, but larger test sets are needed.

MEMSAT-SVM / MEMPACK (2009 + 2012)
- Replaces neural network with support vector machine classifiers.
- Trained using 131 sequences which all have crystal structures available.
- Can additionally identify re-entrant regions.
- Dynamic programming algorithm.
- Pores and pore stoichiometry.

Short peptides bound to globular domains and likely to be unfolded in isolation based on the analysis of interface and accessible surface areas.

Unbound protein domain linkers as annotated in the CATH database.
Crystal structure data sets are predominantly composed of prokaryotic structures.

- Tested using the Möller data set: 79% accuracy
- Tested using the TOPDB data set: 67% accuracy

**PSICOV Workbench (coming soon)**

- PSICOV2, NNPSICOV
  - Residue contact prediction tools
- FILM3
  - De novo Transmembrane Protein Modelling
- ContactTHREADER
  - Fold recognition and modelling
- PPI-PSICOV

**Acknowledgements**

- Domenico Cozzetto
- Federico Minneci
- Tim Nugent
- Dan Buchan

**FFPred 2.0 - Input via PSIPRED server**

**FFPred 2.0 - Input via PSIPRED server**
**5-HT\textsubscript{1B} receptor (Homo sapiens)**

Serotonin receptor 1B:
- serotonin receptor subtype
- well-researched GPCR, with 7 transmembrane helices
- part of serotonin pathways, widely distributed across the central nervous system
- therefore, known signal transduction activity, GPCR activity, ...

**5-HT\textsubscript{1B} receptor – FFpred predictions (I)**

Select "FFPred" tab in results page

GO ontology

High or low reliability of the SVM that was trained to predict the corresponding GO term

Predictions associated to less reliable SVMs are listed at the bottom, on a red background...

...for each category of SVM reliability, annotations are then ranked according to the posterior probability of the prediction being correct

**5-HT\textsubscript{1B} receptor – FFpred predictions (II)**

Complete GO term predictions for the 5-HT\textsubscript{1B} receptor

Expected GO terms for this well characterised protein are predicted with high probability

More predictions may give indications for further experimental investigation

Low-reliability predictions are meant to be further, relatively less safe suggestions

**5-HT\textsubscript{1B} receptor – FFpred predictions (III)**

Used protein features can be found below the GO term predictions

Features like secondary structure (PSIPRED), disorder (DISOPRED), PEST regions, post-translational modifications can be read on the summary diagram or on the detailed amino acid map

Transmembrane topology (MEMSAT-SVM)

Amino acid composition and physico-chemical properties

---

**Human SNW domain-containing protein 1**

SNW1 (UniProt AC Q13573) is a eukaryotic multifunctional protein involved in transcription initiation and repression, mRNA splicing, and cell cycle by interacting with many partners.

The sequence contains high proportion of polar and charged amino acids, typical of disordered proteins.

NMR studies on its role in spliceosome maturation show that the N-terminal 172 positions are disordered and that residues from 59 to 79 fold upon binding the PPL1 protein.
DISOPRED predictions for SNW1

65% of the 172 N-terminal disordered residues correctly classified.

Other predictions agree with common assumptions and consensus data in MobiDB and D2P2.

The disordered protein binding region from position 59 to 79 is predicted with 38% precision.
Genome3D Workshop

Ian Sillitoe, Tony Lewis
UCL

Why Structure?

- Enzyme Active Sites
- Protein-Protein Interfaces

Crystal structure of the Anopheles gambiae 3-hydroxykynurenine transaminase. Rossi et al. PNAS

What is Genome3D?

- Genome3D provides...
  - consensus structural annotations
  - consensus 3D models
- ...by identifying similarities to known protein domains
  - libraries of structural domains from CATH/SCOP
  - domains are classified into “superfamilies”

Domain Structure Classification

- SUPERFAMILY
  - Gene3D
  - PHYRE
  - gDomTHREADER
  - FUGUE

Domain Structure Annotation

- ... Associated functional information from > 10 public sources
- KEGG terms
- GO terms
- Other ...

Genome3D Workflow

- 10 model genomes (Uniprot)
- Pfam representatives
- Identify similarity to “template” domains from CATH / SCOP
- Predict domain boundaries and/or 3D structure
- Map CATH / SCOP superfamilies
- Present data on website (one page per UniProt sequence)

Genome3D Consortium

<table>
<thead>
<tr>
<th>Domain Classification</th>
<th>Principle Investigator</th>
<th>Prediction Type</th>
<th>University</th>
</tr>
</thead>
<tbody>
<tr>
<td>CATH</td>
<td>Orengo</td>
<td>Models</td>
<td>UCL</td>
</tr>
<tr>
<td>SCOP</td>
<td>Marvin</td>
<td>Annotations</td>
<td>Cambridge</td>
</tr>
<tr>
<td>FUGUE</td>
<td>Blundell</td>
<td>Models</td>
<td>UCL</td>
</tr>
<tr>
<td>PhyRE</td>
<td>Gough</td>
<td>Models</td>
<td>UCL</td>
</tr>
<tr>
<td>pDomTHREADER</td>
<td>Bork / Vallely</td>
<td>Annotations</td>
<td>Imperial</td>
</tr>
<tr>
<td>SUPERFAMILY</td>
<td>Gough</td>
<td>Both</td>
<td>SCOP / PDB</td>
</tr>
<tr>
<td>VIVACE</td>
<td>Blundell</td>
<td>Models</td>
<td>Cambridge</td>
</tr>
</tbody>
</table>
CATH / SCOP Mapping

- Find pairs of overlapping SCOP/CATH domains:
  - # residues in CATH domain
  - # residues in SCOP domain
  - # residues in common
- Initially scan and store every overlapping pair (even trivial overlaps)
- Later, reject on the basis of overlap (of smallest domain)

- Find related pairs of SCOP/CATH superfamilies
- Map between
  - SCOP v1.75 (1962 superfamilies)
  - CATH v3.5.0 (2626 superfamilies)
- Assess how well pairs of superfamilies overlap

Gold
763 pairs

Silver
134 pairs

Bronze
532 pairs
SCOP/CATH Mapping

Genome3D: Dataset
Genomes, Pfam, Coverage

Genome3D: Dataset
• 10 model genomes
  – Human (homo sapiens)
  – E. coli (escherichia coli)
  – Baker’s yeast (saccharomyces cerevisiae)
  – Mouse (mus musculus)
  – Mouse-ear cress (arabidopsis thaliana)
  – Fruit fly (drosophila melanogaster)
  – Nematode (caenorhabditis elegans)
  – Malaria parasite (plasmodium falciparum)
  – Bacterium (staphylococcus aureus)
  – Fission yeast (schizosaccharomyces pombe)

Genome3D: Dataset
• Pfam has > 80% coverage of UniProt
  – Exploit Pfam’s coverage
  – Reuse existing pipeline
  – Worked with Pfam to select representative sequences
• Pfam - the 11th Genome

Genome3D: Coverage
Predicted Domain Annotations (Per UniProt Sequence)

Genome3D: Coverage
Predicted 3D Models (Per UniProt Sequence)
Genome3D: Web

www.genome3d.eu

Search

UniProt: Overview

UniProt: Annotations

Predicted Domains
Tutorial after Coffee Break In Foster Court B29