High Throughput Functional Annotation of The Human Proteome

Philip Guthrie[1], Sunil Patel[2], Jennifer Worroll[1], Dana Haley-Vicente[2], John Bramall[1], Jonathan Sheldon[1]

[1]*Confirmant Ltd, Unit 66, The Business Development Centre, Milton Park, Abingdon, Oxon OX14 4RX, UK*
[2]*Accelrys Inc., 9685 Scranton Road, San Diego, CA 92121, USA*

Abstract

A high throughput approach to proteome functional annotation is currently being explored using Confirmant's Protein Atlas of the Human Genome ™ (Protein Atlas) (ref 1) and Accelrys' Discovery Studio (DS) GeneAtlas™ (ref 2) high-throughput functional annotation pipeline. The Protein Atlas contains peptide sequences experimentally derived using high throughput mass spectrometric analysis (MALDI-TOF and MS/MS) of proteins expressed in a wide range of tissues, cell lines, sub-cellular fractions and disease states. This data is then mapped back to the chromosomal backbone, providing a protein-centric view of the genome. A set of 14,000 Protein Atlas sequences, including many novel transcripts of proteins of unknown function, was run through the DS GeneAtlas pipeline to assign putative function through sequence similarity detection, homology modeling, and fold recognition methods. Using template searching, DS GeneAtlas searches for relationships between query sequences and known protein structures, motifs, and folds (ref 3). Subsequent inferences and assignment of the target protein's function is based on its homology to the experimentally derived template protein and the models generated as part of the pipeline. The functional annotation created by DS GeneAtlas was then stored in DS AtlasStore to be queried and analysed.

Focusing on possible drug targets, tyrosine kinase was used as a model system to assess the quality of the annotation. The keywords 'tyrosine kinase' were used to query DS AtlasStore™ (ref 2) and the results were filtered to identify potentially novel candidate tyrosine kinases that were further investigated by examination of structural characteristic features of tyrosine kinases. These features included the presence of suitable binding pockets, the appropriate spatial distribution of significant residues within conserved domains, and more. Next, we investigated any known tyrosine kinases for which the Protein Atlas demonstrated genes with peptide-validated novel exons or alternative transcripts, since these features could have significant bearing on the mode of action of the expressed protein. The results of this analysis will be presented along with examples of fully annotated novel tyrosine kinases.

Introduction

Confirmant's Protein Atlas (ref 1) contains experimentally derived protein sequences from over 14,000 genes. Many of these sequences are not in the public domain either because they are from a previously unidentified gene or they are novel splice variants. In an effort to characterise both the novel proteins and those proteins that are sparsely/poorly annotated in the public domain, we ran the Protein Atlas derived sequences through Accelrys' DS GeneAtlas pipeline. The process and results are described in this poster.

Methodology

The Protein Atlas (ref 1) contains peptide sequences experimentally derived using high throughput mass spectrometric analysis (MALDI-TOF and MS/MS) of proteins expressed in a wide range of human tissues, cell lines, and body fluids from a range of disease states. A "virtual transcriptome" is constructed using public domain sequences, gene predictions and potential alternative splice sites against which the MS/MS spectra are searched, using Sequest (ref 4). The resulting peptide sequences, along with MS masses are mapped back to the human genome, utilising the knowledge that peptides from the same protein must map back to the same gene. The mapped peptides are used to deduce both the exon structure of the gene and the protein forms seen in each sample. Figure 1 (below) shows an overview of this process.

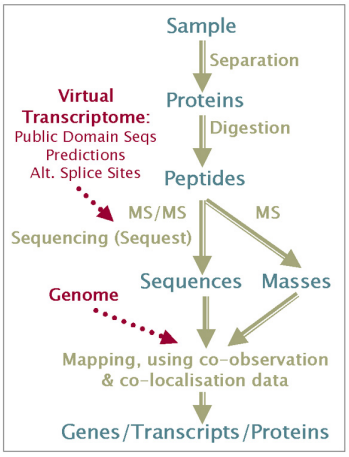

Figure 1: Overview of Confirmant's pipeline, showing how proteomic data is obtained and analyzed to determine the protein variants present in each individual sample.

Over 14,000 protein sequences from Confirmant's Protein Atlas (ref 1) were run through Accelrys' DS GeneAtlas™ (ref 2, 3), an automated protein annotation pipeline for analysing protein sequences and identifying their biochemical function. The DS GeneAtlas pipeline automates and integrates several steps into one seamless operation, collapsing the genomic information explosion and converting it into information and knowledge. This automated pipeline populates the DS AtlasStore™ database with 3D structure prediction and functional annotation of the sequences. DS AtlasStore is an Oracle relational database schema, designed to store sequence data, family information, output from DS GeneAtlas and other annotations. DS AtlasStore provides a graphical interface that allows visualisation of both sequence and structural annotations, implemented as a desktop application.

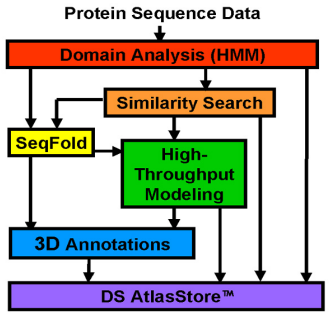

Figure 2: Schematic representation of the DS GeneAtlas™ pipeline, a high throughput pipeline for functional annotation of protein sequences. The resulting annotations are stored into DS AtlasStore™.

The major components of DS GeneAtlas are shown in figure 2. The pipeline consists of identification of functional domains of protein sequences, homology searching using PSI-BLAST, fold recognition using SeqFold (ref 5) High throughput homology modelling (HTM) using MODELER (ref 6), and function annotation using 3D motif searches (ref 7). Based on the observation by J.Park, et al (ref 8), each sequence similarity searching method gives a unique set of hits which are different from other methods. Therefore, the DS GeneAtlas pipeline uses as many methods as possible to extend the homology recognition between query sequence and sequences with known structure or function to maximise the function assignment to the query sequence. Using this strategy in the pipeline increases the confidence of the assignment when a query sequence has consensus hits by different methods. The quality of models generated is assessed against Profiles-3D (Verify) (ref 9) scores and Potential Mean Force (PMF) scores.

From the full set of annotated Confirmant sequences those with a tyrosine kinase function were selected as follows. Firstly a shortlist of all possible tyrosine kinases was produced by searching the DS GeneAtlas annotations using simple text search of the annotations, looking for 'tyrosine AND kinase'. The alignments for each sequence were then considered. Sequences with short alignments and/or poor E-values were rejected and the models for the remaining sequences were inspected. Of these models 82 were considered to be likely tyrosine kinases, based on the sequence identity/similarity of the alignments, high verify scores, high PMF scores, and good conservation of the binding pocket residues between Confirmant sequence and template sequence (see Table 1).

The SwissProt (ref 10) and Sugen (ref 11) datasets were searched for each of the 82 sequences and for the sequences or record that were found, the annotations were compared with those produced by DS GeneAtlas. (A Swissprot/Sugen record was considered to match a Confirmant sequence if the only sequence differences could be explained in terms of novel exons/splice variants in the Confirmant data set).

Results

% Seq similarity	Modeling method	No. selected	E values
> 90	HTM	10	$1.0e^{-23}$ - 0.0
	PSIBLAST	3	$2.6e^{-11}$ - 5.2^{-36}
	SEQFOLD	7	$2.6e^{-39}$ - 0.0
70-90	HTM	4	$2.1e^{-31}$ - $1.6e^{-99}$
	PSIBLAST	0	...
	SEQFOLD	27	$1.6e^{-25}$ - $1.1e^{-99}$
50-70	HTM	6	$4.8e^{-26}$ - $3.2e^{-88}$
	PSIBLAST	5	$6.4e^{-14}$ - $1.1e^{-85}$
	SEQFOLD	3	$9.6e^{-67}$ - $1.6e^{-81}$
30-50	HTM	6	$1.4e^{-4}$ - $9.6e^{-39}$
	PSIBLAST	11	$1.8e^{-2}$ - $5.2e^{-81}$
	SEQFOLD	0	...
		82 total	

Table 1: Breakdown of models selected as potential tyrosine kinases, by % sequence similarity

Seqs.	SwissProt Annotation	Sugen Annotation[3]
5	no SwissProt entry	N/A
2	no SwissProt entry	Other Kinase
2	no SwissProt entry	Tyrosine Kinase
12	Non - Kinase	N/A
2	Other Kinase	N/A
6	Other Kinase	Other Kinase
1	Other Kinase	Tyrosine Kinase like
2	Tyrosine Kinase	N/A
1	Tyrosine Kinase	Other Kinase
43	Tyrosine Kinase	Tyrosine Kinase
1	Tyrosine Kinase like	Other Kinase
5	Tyrosine Kinase like	Tyrosine Kinase like

Table 2: Public domain annotations for the 82 selected sequences, annotated by the DS GeneAtlas™ pipeline as Tyrosine Kinases.

Of the 82 potential tyrosine kinases identified at a first pass with high stringency (table 1) only 43 had shown to be annotated as tyrosine kinases by both Sugen and SwissProt (table 2). Examples of models for proteins CFM011508 and CFM011666, which were previously not annotated as tyrosine kinase by either SwissProt or Sugen are show in tables 3 & 4, with a ribbon diagram of the model for the latter sequence superimposed on its template shown in Figure 3.

Sequence ID	CFM011508
Model ID	CFM011508_68
Alignment size	93
Model score	1.0
Template	TFAM000109
% Sequence Identity	52.04
% Sequence similarity	65.6
E value	$3.2e^{-28}$
Verify score	1.02
PMF score	1.0
BP residues conserved	2/14
Description: P56LCK TYROSINE KINASE; PDB template1blj	

Table 3: Attributes of CFM011508: a potential novel tyrosine kinase.

Sequence ID	CFM011666
Model ID	CFM011666_134
Alignment size	259
Model score	0.90
Template	1fgkB
% Sequence identity	33.06
% Sequence similarity	54.4
E value	$2.3e^{-84}$
Verify score	0.42
PMF score	1.0
BP residues conserved	16/34
Description: HUMAN FIBROBLAST GROWTH FACTOR RECEPTOR WITH TYROSINE KINASE DOMAIN PDB template1fgkB	

Table 4: Attributes of CFM011666: a potential novel tyrosine kinase.

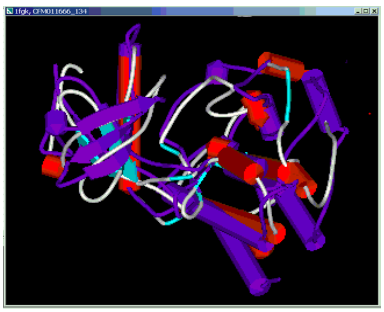

Figure 3: Sequence CFM011666 superimposed on template structure 1fgkB (purple) (see table 4, above).

Fourteen of the 82 sequences had exons identified by Confirmant which were not present in the corresponding Ensembl gene (ref 12) (figure 4). Some of these exons interrupt the domains identified by DS GeneAtlas and so it can be inferred that they affect the structure and therefore the function of the protein. This is useful information in itself as it provides clues as to which function(s) of the protein may be altered/disrupted, but unfortunately the resolution of the models is not yet high enough to predict the effects of the novel exons with confidence.

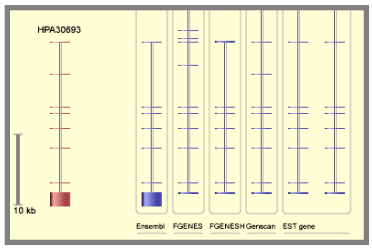

Figure 4: Diagram showing novel 2nd exon in Confirmant gene (red) aligned with Ensembl gene & other predicted genes. (diagram taken from Protein Atlas GUI).
Note: FGenes, FGenesH and EST predictions extend upstream (not shown).

Conclusion

As the number of solved biological structures increases and the methods of DS GeneAtlas are further refined, we can expect to see novel functions assigned to many known proteins, and the number of members in protein families is likely to increase, even for well studied families such as the tyrosine kinases. The effect of modelling can be expected to be even more dramatic for novel sequences such as the novel splice variant and products of novel genes discovered by Confirmant and defined in the Protein Atlas: the functional annotation generated by the synergistic approach of DS Modeling applied to HPA sequences will provide valuable information for characterising the roles of these novel proteins in disease and health.

References

1. Protein Atlas of the Human Genome™ www.confirmant.com
2. DS Modeling 1.1 (DS GeneAtlas™ and DS AtlasStore™, Accelrys, Inc., (2003).
3. Kitson, D., et al. Briefings in Bioinformatics. **3** (2002) 1-13.
4. Sequest (Copyright 1993-96, Molecular Biotechnology, Univ. of Washington, J. Eng/J. Yates Licensed to Finnigan MAT)
5. Olszewski, K.A., et al. Theor. Chem. Acc. **11**, (1999) 57.
6. Sali, A. & Blundell, T.L., Mol. Biol., **234**(1993a) 779-815.
7. Milik. M, et al., R. Guigo and D. Gusfield (Eds.), Springer Verlag, Berlin, Lecture Notes in Computer Science **2452** (2002) 172-184.
8. Hubbard T. J, & Park J. Proteins. **23** (1995) 398-402.
9. Bowie, J. U., et al. Science **253** (1991) 164-170.
10. Boekmann, B. et al, Nucleic Acid Res **31** (2003) 365-370
11. G. Manning et al. Science **298** (2002) 1912-1934
12. Hubbard T., et al Nucleic Acid Res **30** (2002) 38-41