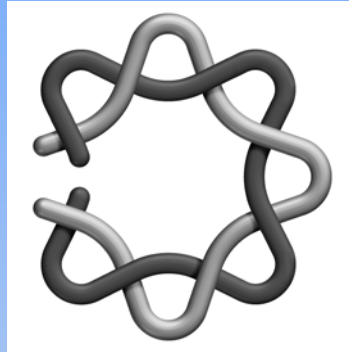


# Practical Experience with Grid Computing



Liam J. McGuffin

Head of Bioinformatics and Systems Biology Unit,  
The BioCentre,  
University of Reading

[l.j.mcguffin@reading.ac.uk](mailto:l.j.mcguffin@reading.ac.uk)



The University of Reading

# e-PROTEIN

<http://www.e-protein.org>

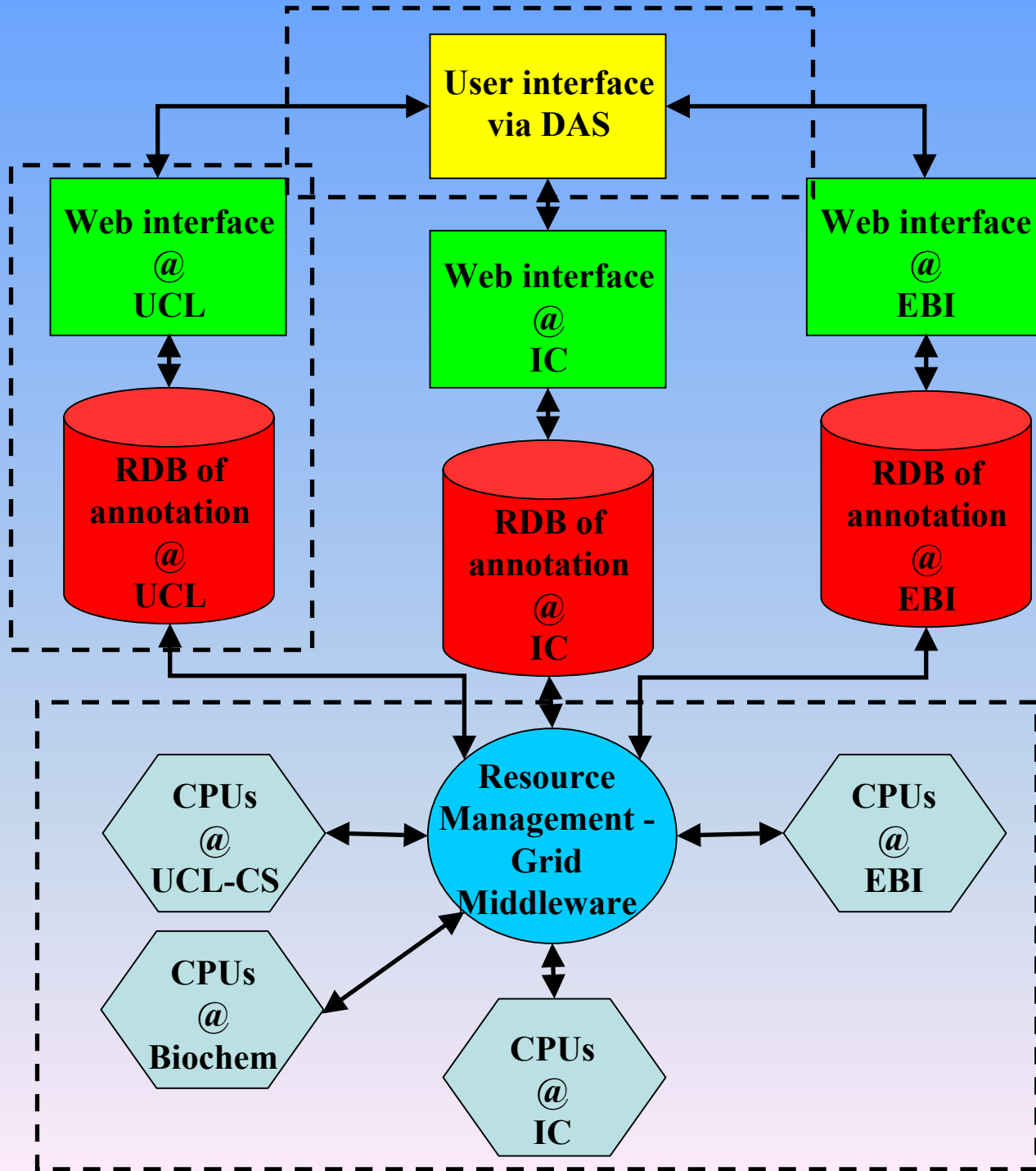
“To provide a fully automated distributed pipeline for large-scale structural and functional annotation of all major proteomes via the use of cutting-edge computer Grid technologies.”

- University College London
- Imperial College, London
- European Bioinformatics Institute, Cambridge

- Integrate databases using the Protein Distributed Annotation system (DAS)

- Deposit structural annotations in relational databases with web interfaces: Genomic Threading Database (GTD) at UCL

- Harness power of many computer clusters at multiple sites UCL and Imperial and EBI



# The Genomic Threading Database

- <http://bioinf.cs.ucl.ac.uk/GTD>
- The GTD contains structural annotations of proteomes from key organisms
- GenTHREADER/mGenTHREADER fold recognition methods are currently used for structural annotations
- Annotation jobs are distributed across clusters of computers, currently at UCL and Imperial college, using Grid technology

# GTD data 1

**The Genomic Threading Database presently contains:**

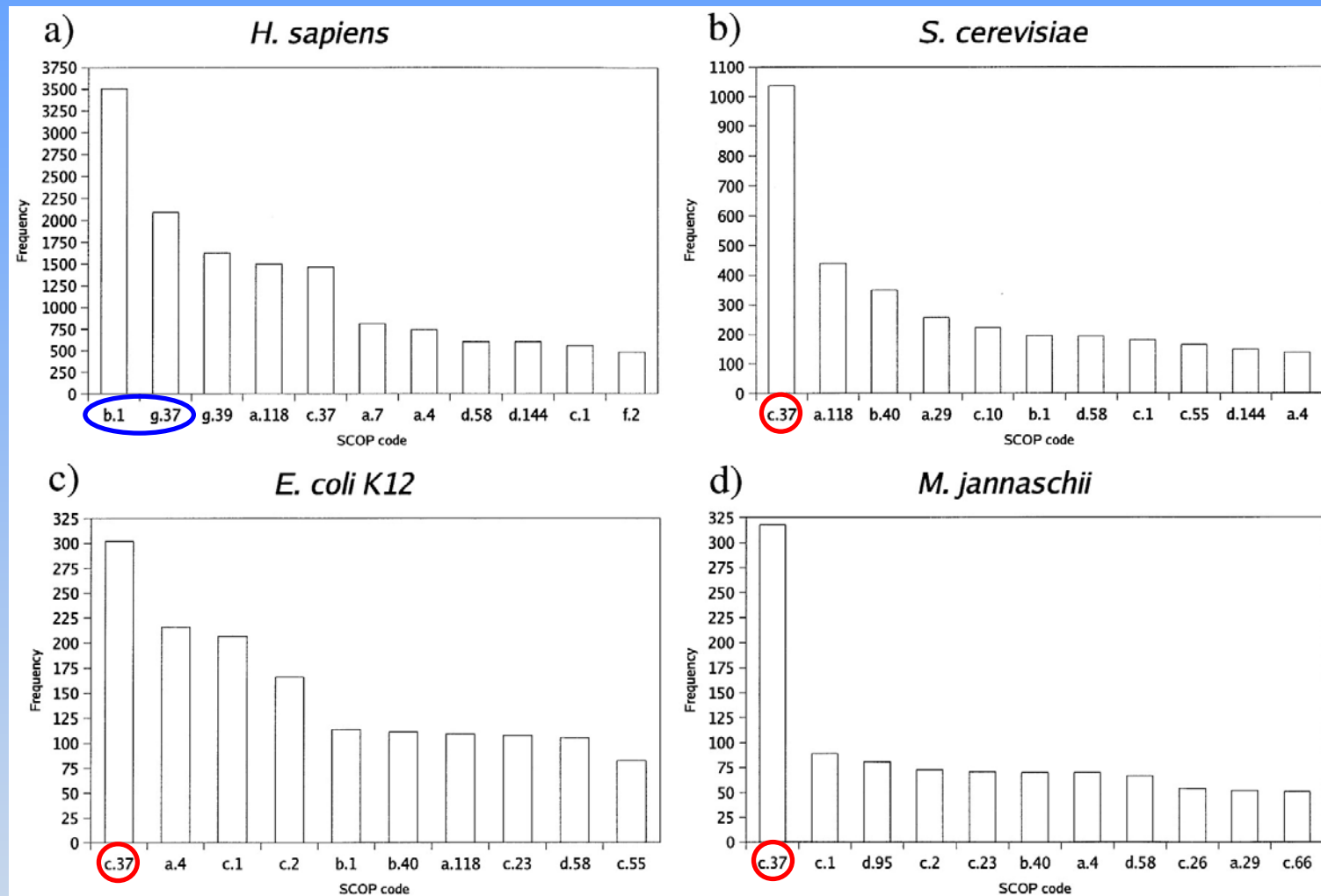
- **261 annotated genomes**
- **1,219,063 annotated sequences**
- **265,673,588 aligned residues**
- **On average > 80% of globular proteins have assigned folds with  $p < 0.01$**

**<http://bioinf.cs.ucl.ac.uk/GTD>**

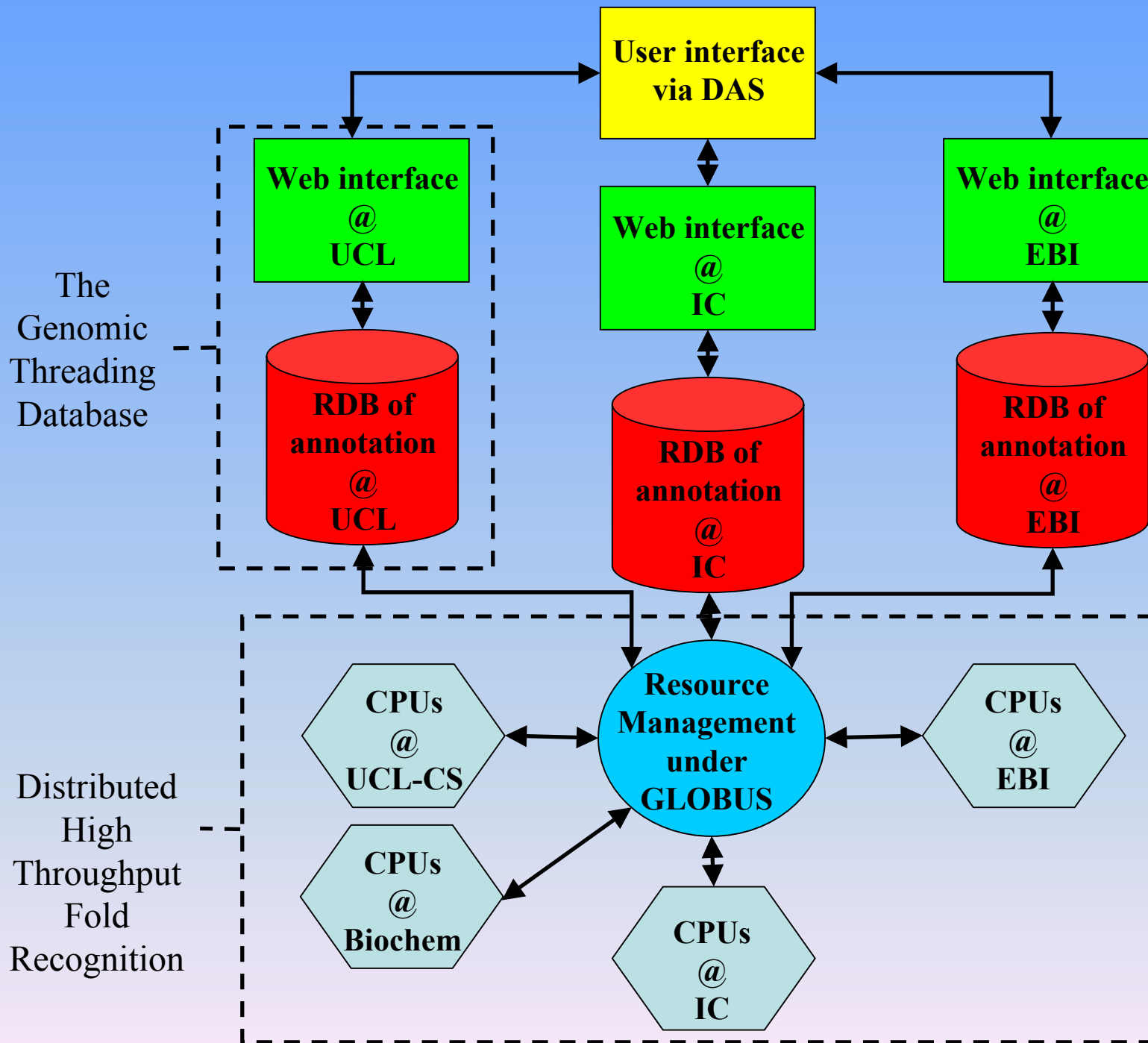
## **Publications:**

1. **McGuffin, L. J.,** Street, S., Bryson K., Sorensen, S. A. & Jones, D. T. (2004) The Genomic Threading Database: a comprehensive resource for structural annotations of the genomes from key organisms. *Nucleic Acids Res.*, 32, D196-D199.
2. **McGuffin, L. J.,** Street, S., Sorensen, S. A. & Jones, D. T. (2004) The Genomic Threading Database. *Bioinformatics*, 20, 131-132.

# GTD data 2



Each SCOP code relates to the following folding types: a.4, DNA/RNA-binding 3-helical bundle; a.7, spectrin repeat-like; a.29, bromodomain-like; a.118, - superhelix; b.1, immunoglobulin-like sandwich; b.40 OB-fold; c.1, TIM barrel; c.2, NAD(p)-binding Rossmann-fold domains; c.10, leucine-rich repeat; c.23, flavodoxin-like; c.26, adenine nucleotide hydrolase-like; c.37, P-loop-containing nucleotide triphosphate hydrolases; c.55, ribonuclease H-like motif; c.66, S-adenosul-L-methionine-dependent methyltransferases; d.58, ferredoxin-like; d.95, homing endonuclease-like; d.144, protein kinase-like (PK-like); f.2, membrane all-; g.37, C2H2 and C2HC zinc fingers; g.39, glucocorticoid receptor-like (DNA-binding domain).





# Bioinformatics Unit

[GTD home>](#)

## JPortal Registration

[Info](#)

This form allows you to register with JPortal which allows you to submit proteome sequences for annotation. Further information and references can be found on the [Genomic Threading Database home page](#).

[Registration](#)Name: Affiliation: Department: Project: Email: [» contact](#)

McGuffin, L. J., Street, S., Bryson K., Sorensen, S. A. & Jones, D. T. (2004) The Genomic Threading Database: a comprehensive resource for structural annotations of the genomes from key organisms. *Nucleic Acids Res.*, 32, D196-D199.



# Bioinformatics Unit

[GTD home>](#)

## JPortal Login

[Info](#)

This form allows you to login to JPortal where you may submit proteome sequences for annotation. Further information and references can be found on the [Genomic Threading Database home page](#).

[Login](#)Username: Password: [» contact](#)

McGuffin, L. J., Street, S., Bryson K., Sorensen, S. A. & Jones, D. T. (2004) The Genomic Threading Database: a comprehensive resource for structural annotations of the genomes from key organisms. *Nucleic Acids Res.*, 32, D196-D199.

[UCL home](#) [Bioinformatics home](#) [GTD home](#) [McGuffin home](#)



# Bioinformatics Unit

## Preparing sequences for annotation

Uploading proteome	COMPLETE
Assigning unique sequence identifiers	COMPLETE
Filtering sequences	COMPLETE
Dividing proteome file for distribution across clusters	COMPLETE

Your sequences have been submitted to the following clusters for annotation:

- cs\_G\_theta\_15\_19\_18\_292\_22-4-2004\_cs.1.fasta.md5.pfill
- cs\_G\_theta\_15\_19\_18\_292\_22-4-2004\_cs.2.fasta.md5.pfill
- cs\_G\_theta\_15\_19\_18\_292\_22-4-2004\_cs.3.fasta.md5.pfill
- cs\_G\_theta\_15\_19\_18\_292\_22-4-2004\_cs.4.fasta.md5.pfill
- cs\_G\_theta\_15\_19\_18\_292\_22-4-2004\_cs.5.fasta.md5.pfill
- cs\_G\_theta\_15\_19\_18\_292\_22-4-2004\_cs.6.fasta.md5.pfill
- cs\_G\_theta\_15\_19\_18\_292\_22-4-2004\_cs.7.fasta.md5.pfill
- cs\_G\_theta\_15\_19\_18\_292\_22-4-2004\_cs.8.fasta.md5.pfill
- cs\_G\_theta\_15\_19\_18\_292\_22-4-2004\_cs.9.fasta.md5.pfill
- cs\_G\_theta\_15\_19\_18\_292\_22-4-2004\_cs.10.fasta.md5.pfill
- cs\_G\_theta\_15\_19\_18\_292\_22-4-2004\_cs.11.fasta.md5.pfill
- cs\_G\_theta\_15\_19\_18\_292\_22-4-2004\_cs.12.fasta.md5.pfill
- cs\_G\_theta\_15\_19\_18\_292\_22-4-2004\_cs.13.fasta.md5.pfill
- cs\_G\_theta\_15\_19\_18\_292\_22-4-2004\_cs.14.fasta.md5.pfill
- cs\_G\_theta\_15\_19\_18\_292\_22-4-2004\_cs.15.fasta.md5.pfill
- cs\_G\_theta\_15\_19\_18\_292\_22-4-2004\_cs.16.fasta.md5.pfill
- cs\_G\_theta\_15\_19\_18\_292\_22-4-2004\_cs.17.fasta.md5.pfill
- cs\_G\_theta\_15\_19\_18\_292\_22-4-2004\_cs.18.fasta.md5.pfill
- cs\_G\_theta\_15\_19\_18\_292\_22-4-2004\_cs.19.fasta.md5.pfill
- cs\_G\_theta\_15\_19\_18\_292\_22-4-2004\_cs.20.fasta.md5.pfill
- cs\_G\_theta\_15\_19\_18\_292\_22-4-2004\_cs.21.fasta.md5.pfill
- cs\_G\_theta\_15\_19\_18\_292\_22-4-2004\_cs.22.fasta.md5.pfill
- cs\_G\_theta\_15\_19\_18\_292\_22-4-2004\_cs.23.fasta.md5.pfill
- cs\_G\_theta\_15\_19\_18\_292\_22-4-2004\_cs.24.fasta.md5.pfill
- cs\_G\_theta\_15\_19\_18\_292\_22-4-2004\_cs.25.fasta.md5.pfill
- cs\_G\_theta\_15\_19\_18\_292\_22-4-2004\_cs.26.fasta.md5.pfill
- cs\_G\_theta\_15\_19\_18\_292\_22-4-2004\_cs.27.fasta.md5.pfill
- cs\_G\_theta\_15\_19\_18\_292\_22-4-2004\_cs.28.fasta.md5.pfill
- cs\_G\_theta\_15\_19\_18\_292\_22-4-2004\_cs.29.fasta.md5.pfill
- cs\_G\_theta\_15\_19\_18\_292\_22-4-2004\_cs.30.fasta.md5.pfill
- cs\_G\_theta\_15\_19\_18\_292\_22-4-2004\_cs.31.fasta.md5.pfill
- cs\_G\_theta\_15\_19\_18\_292\_22-4-2004\_cs.32.fasta.md5.pfill
- cs\_G\_theta\_15\_19\_18\_292\_22-4-2004\_cs.33.fasta.md5.pfill
- cs\_G\_theta\_15\_19\_18\_292\_22-4-2004\_cs.34.fasta.md5.pfill
- cs\_G\_theta\_15\_19\_18\_292\_22-4-2004\_cs.35.fasta.md5.pfill
- cs\_G\_theta\_15\_19\_18\_292\_22-4-2004\_cs.36.fasta.md5.pfill
- cs\_G\_theta\_15\_19\_18\_292\_22-4-2004\_cs.37.fasta.md5.pfill



# Bioinformatics Unit

## Progress of mGenTHREADER annotation for G\_theta\_11\_20\_42\_115\_22-4-2004

Total number of sequences	Total Sequences Complete	Total Sequences Failed
466	466	0

### Breakdown of individual jobs

Cluster	Node	Job Number	Job Size (sequences)	Sequences Complete	Sequences Failed
biochem	gere69	1	1	1	0
biochem	gere75	10	1	1	0
biochem	gere71	100	1	1	0
biochem	gere73	101	1	1	0
biochem	gere77	102	1	1	0
biochem	gere69	103	1	1	0
biochem	gere69	104	1	1	0
biochem	gere67	105	1	1	0
biochem	gere70	106	1	1	0
biochem	gere75	107	1	1	0
biochem	gere75	108	1	1	0
biochem	gere66	109	1	1	0
biochem	gere67	11	1	1	0
biochem	gere74	110	1	1	0
biochem	gere77	111	1	1	0
biochem	gere71	112	1	1	0
biochem	gere66	113	1	1	0
biochem	gere67	114	1	1	0
biochem	gere68	115	1	1	0
biochem	gere73	116	1	1	0
biochem	gere74	117	1	1	0
biochem	gere73	118	1	1	0
biochem	gere68	119	1	1	0
biochem	gere73	12	1	1	0
biochem	gere77	120	1	1	0
biochem	gere67	121	1	1	0
biochem	gere69	122	1	1	0
biochem	gere71	123	1	1	0
biochem	gere70	124	1	1	0



# Bioinformatics Unit

[McGuffin home>](#)[Jones home>](#)

## The Genomic Threading Database

Liam J. McGuffin &amp; David T. Jones

The Genomic Threading Database (GTD) contains structural annotations of proteomes, translated from the genomes of key organisms. Annotations are made using a modified version of our recently developed GenTHREADER software.

The GTD is part of the [e-Protein](#) project. [More...](#)

Number of annotated genomes: **174**

Number of annotated sequences: **812,568**

Number of aligned residues: **171,074,491**

For queries regarding the GTD: [l.mcguffin@cs.ucl.ac.uk](mailto:l.mcguffin@cs.ucl.ac.uk)

### Options:

- [Search the database](#)
- [Summary of predictions](#)
- [Download GTD lists](#)

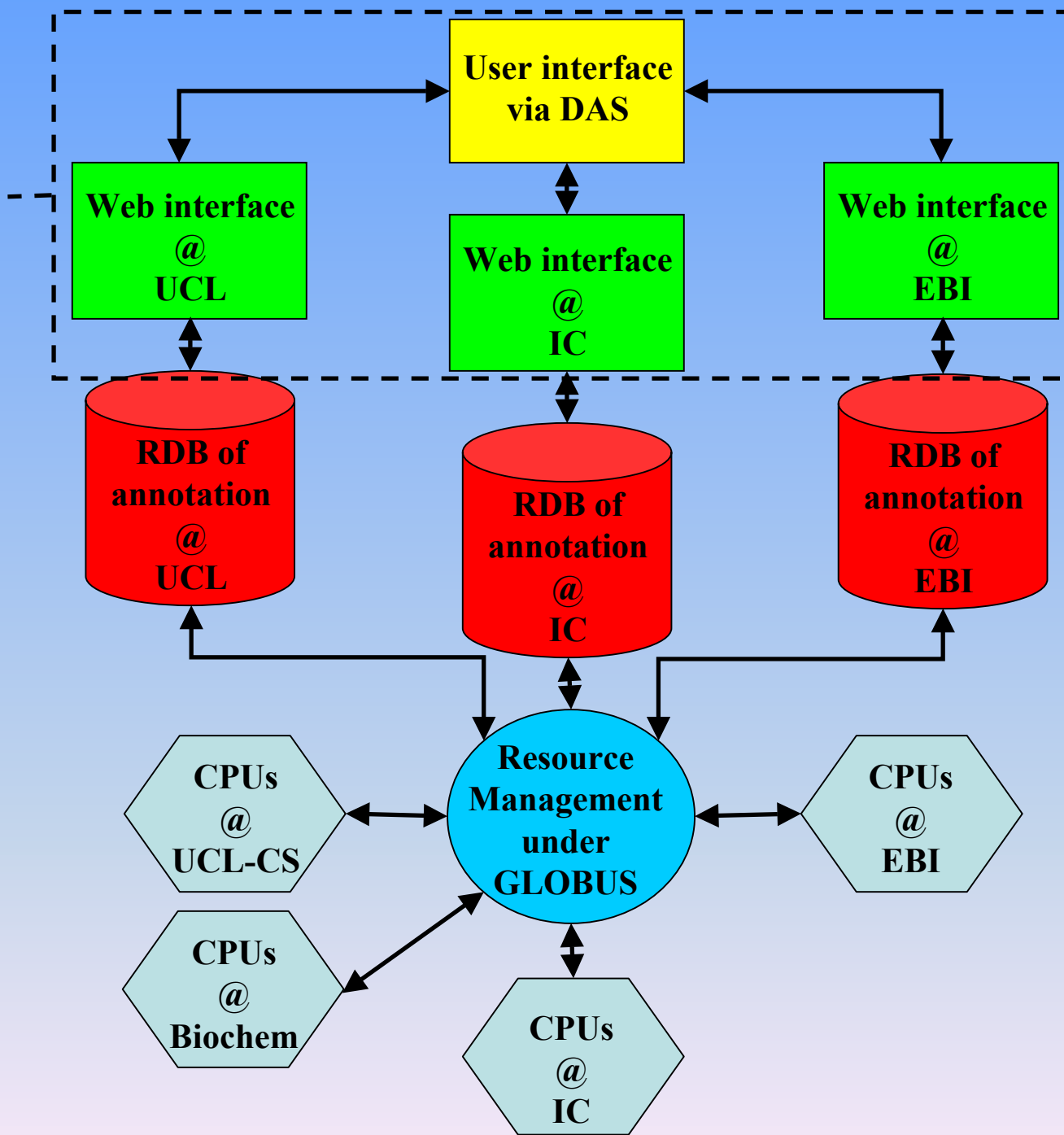
Please cite the following references:

*The GTD*

## Description

[Genomic Threading Database Help Page](#)

Protein  
DAS for  
integration  
of structural  
and  
functional  
data



**Ensembl Home**[EBI Home](#)[Sanger Home](#)[Trace Server](#)[Genome Central](#)[IPI](#)**Search****Help****Documentation****News****BLAST****SSAH****Download****Sitemap****Disclaimer****Ensembl ProteinDAS and GeneDAS****Overview**

The Distributed Annotation System ([DAS](#)) has been widely used by Ensembl to include external annotations, including user's own data, on ContigView displays.

**ProteinDAS**

ProteinDAS uses the DAS protocol to exchange protein annotation. In this case, SwissProt peptide sequence are used as the common reference.

Ensembl currently provides two proof-of-concept ProteinDAS annotation servers:

- \* <http://das.ensembl.org/das/swissprot> serving SwissProt annotations, and
- \* <http://das.ensembl.org/das/interpro> serving InterPro annotations.

Example requests would be:

- \* [http://das.ensembl.org/das/swissprot/features?segment=MIP\\_MOUSE](http://das.ensembl.org/das/swissprot/features?segment=MIP_MOUSE), and
- \* [http://das.ensembl.org/das/interpro/features?segment=CLAT\\_HUMAN](http://das.ensembl.org/das/interpro/features?segment=CLAT_HUMAN)

ProteinDAS annotations can be browsed from Ensembl ProtView pages, e.g:

- \* [http://www.ensembl.org/Mus\\_musculus/protview?peptide=MIP\\_MOUSE](http://www.ensembl.org/Mus_musculus/protview?peptide=MIP_MOUSE), and
- \* [http://www.ensembl.org/Homo\\_sapiens/protview?peptide=CLAT\\_HUMAN](http://www.ensembl.org/Homo_sapiens/protview?peptide=CLAT_HUMAN),

The "Manage Sources" button accesses a script (DasConfView) that allows users to add/edit new DAS sources. Please note that most of the available sources use standard 'genomic' reference sequences.

For further info on ProteinDAS, or for details on sharing protein annotation, please [get in touch](#)

**GeneDAS**

GeneDAS is a semantic extension to the DAS protocol that allows the exchange of annotations via reference identifiers rather than reference sequence. Location data is irrelevant, and annotation applies to the entire entity referenced by the ID.

Ensembl currently uses the SwissProt ProteinDAS server (above) in a GeneDAS capacity. GeneDAS annotations are defined as having identical 'id' attributes for both 'SEGMENT' and 'FEATURE' tags. An example DAS request with both GeneDAS and ProteinDAS annotations is:

- \* [http://das.ensembl.org/das/swissprot/features?segment=MIP\\_MOUSE](http://das.ensembl.org/das/swissprot/features?segment=MIP_MOUSE)

The corresponding Ensembl GeneView page is:

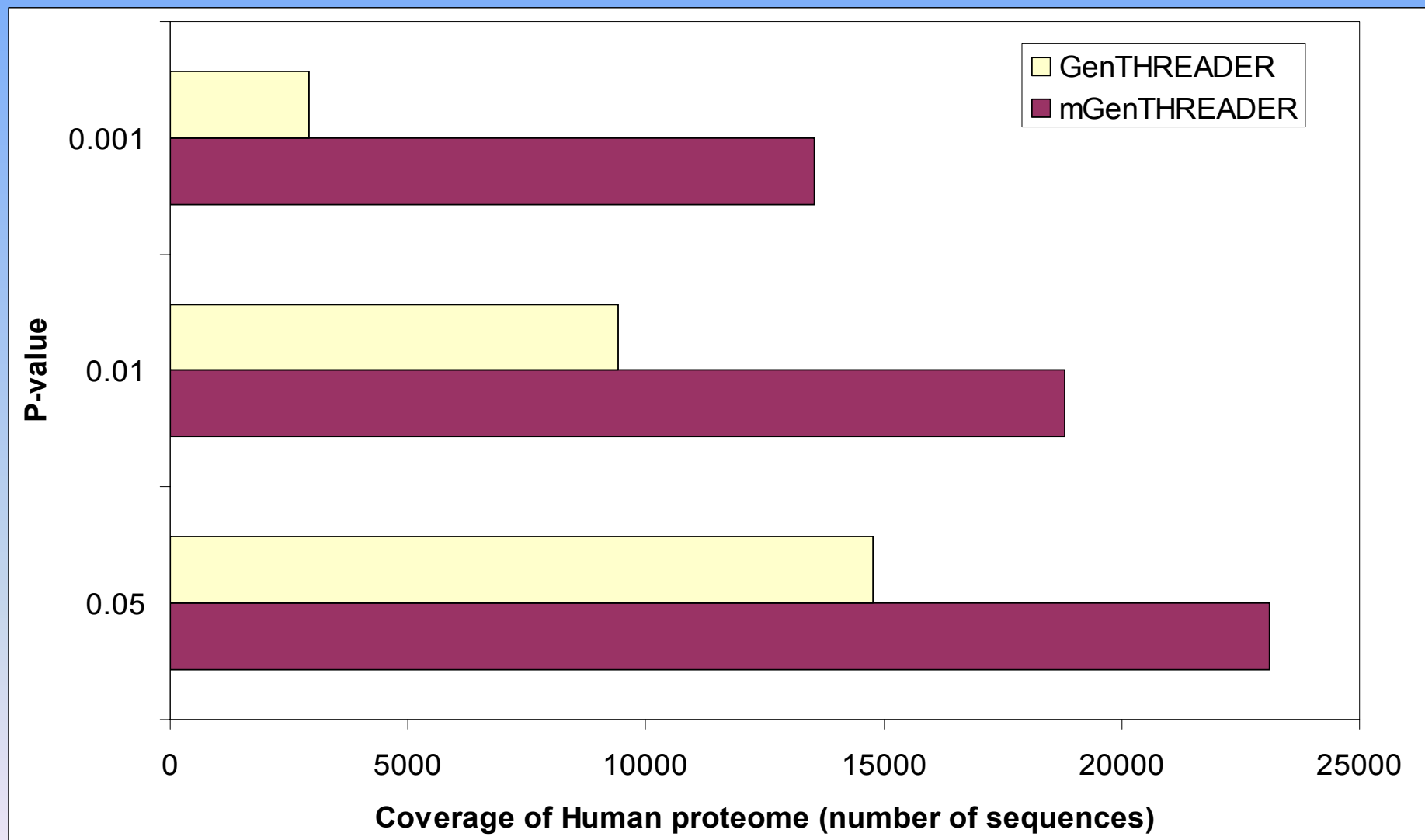
- \* [http://www.ensembl.org/Mus\\_musculus/geneview?gene=MIP\\_MOUSE](http://www.ensembl.org/Mus_musculus/geneview?gene=MIP_MOUSE)

# Human proteome annotation in 24 hours

- Profile-profile version of mGenTHREADER
- Grid Middleware - JYDE (Job Yield Distribution Environment)
- 500+ CPUs, 3 independent Grid domains
  - Imperial LeSC
  - UCL CCC
  - UCL CS
- 99.9% of Human proteome annotated in under 24h

# Coverage

## mGenTHREADER v GenTHREADER (Profile-Profile V Sequence-Profile)



# CPU Time

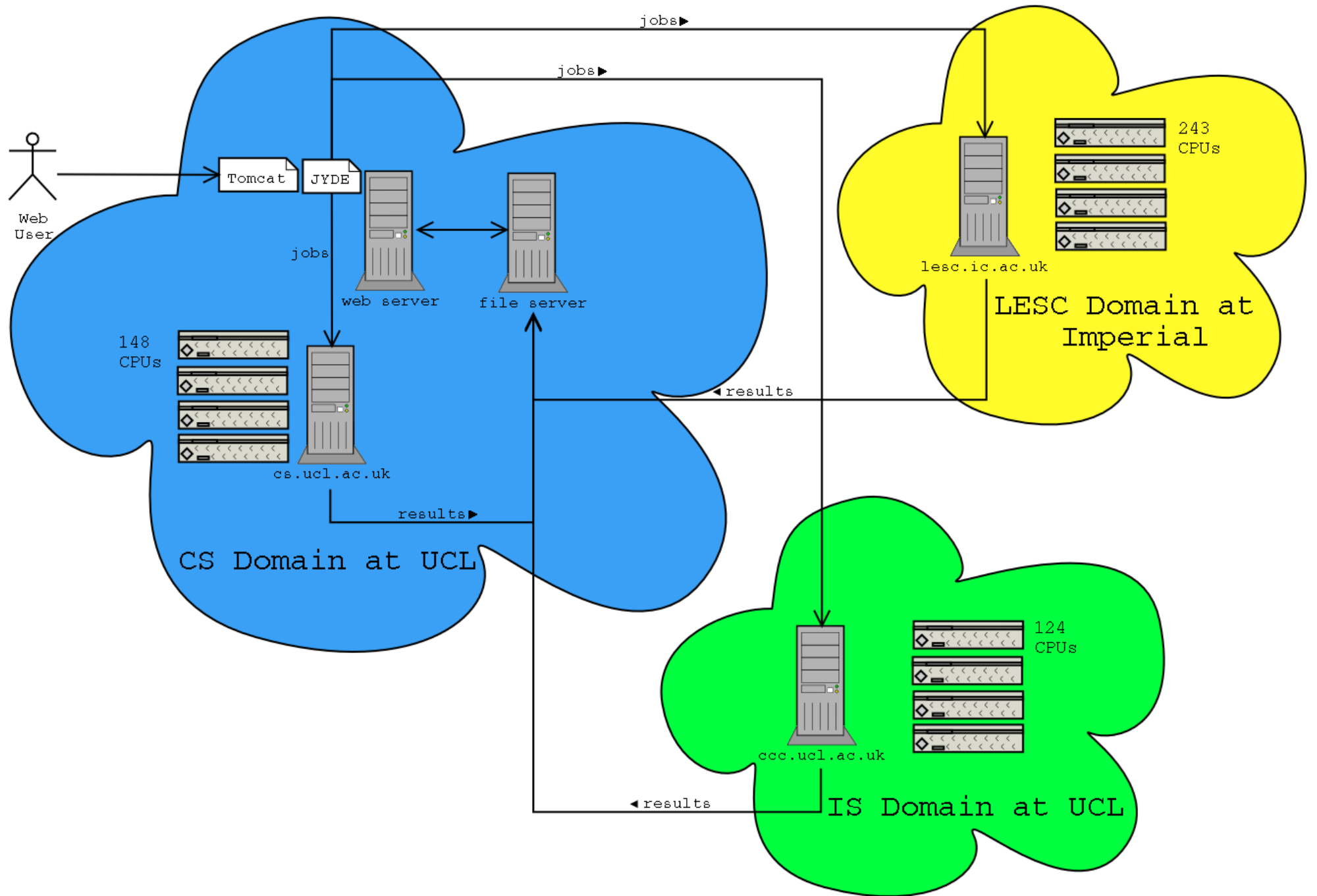
## mGenTHREADER v GenTHREADER

### (Profile-Profile V Sequence-Profile)

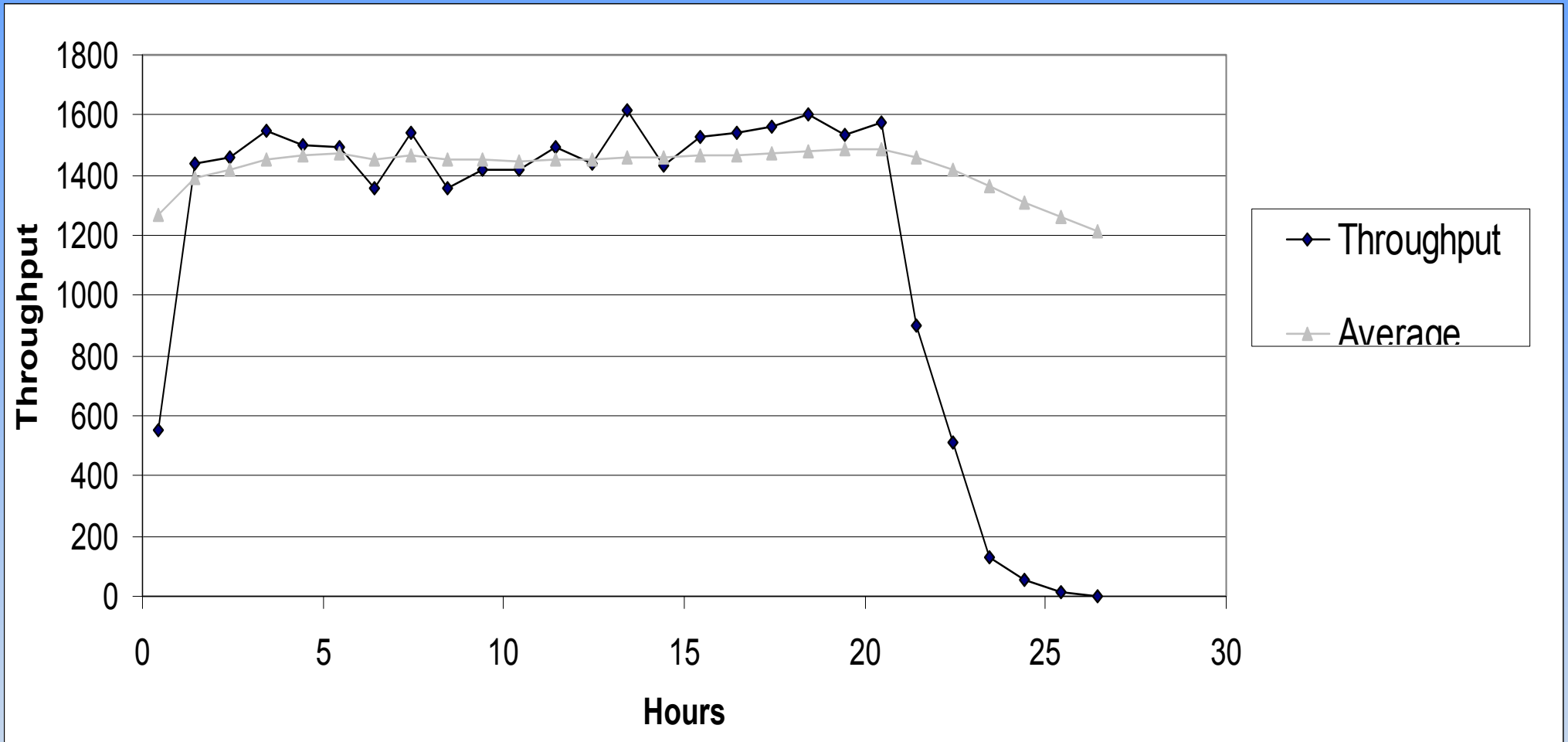
	GenTHREADER	mGenTHREADER
Mean sequences per hour on 148 identical CPUs	1236	230
Mean sequences per hour on 1 CPU	8.35	1.55
Mean time for 1 Human sequence on 1 CPU (min)	7.18	38.61
Estimated time for all Human sequences on 1 CPU (days)	159.70	858.24

- GenTHREADER is 5.4 times the speed of mGenTHREADER
- If we use 1 CPU to annotate Human proteome:
  - GenTHREADER would take < 0.5 years
  - mGenTHREADER would take > 2.4 years

# JYDE pipeline

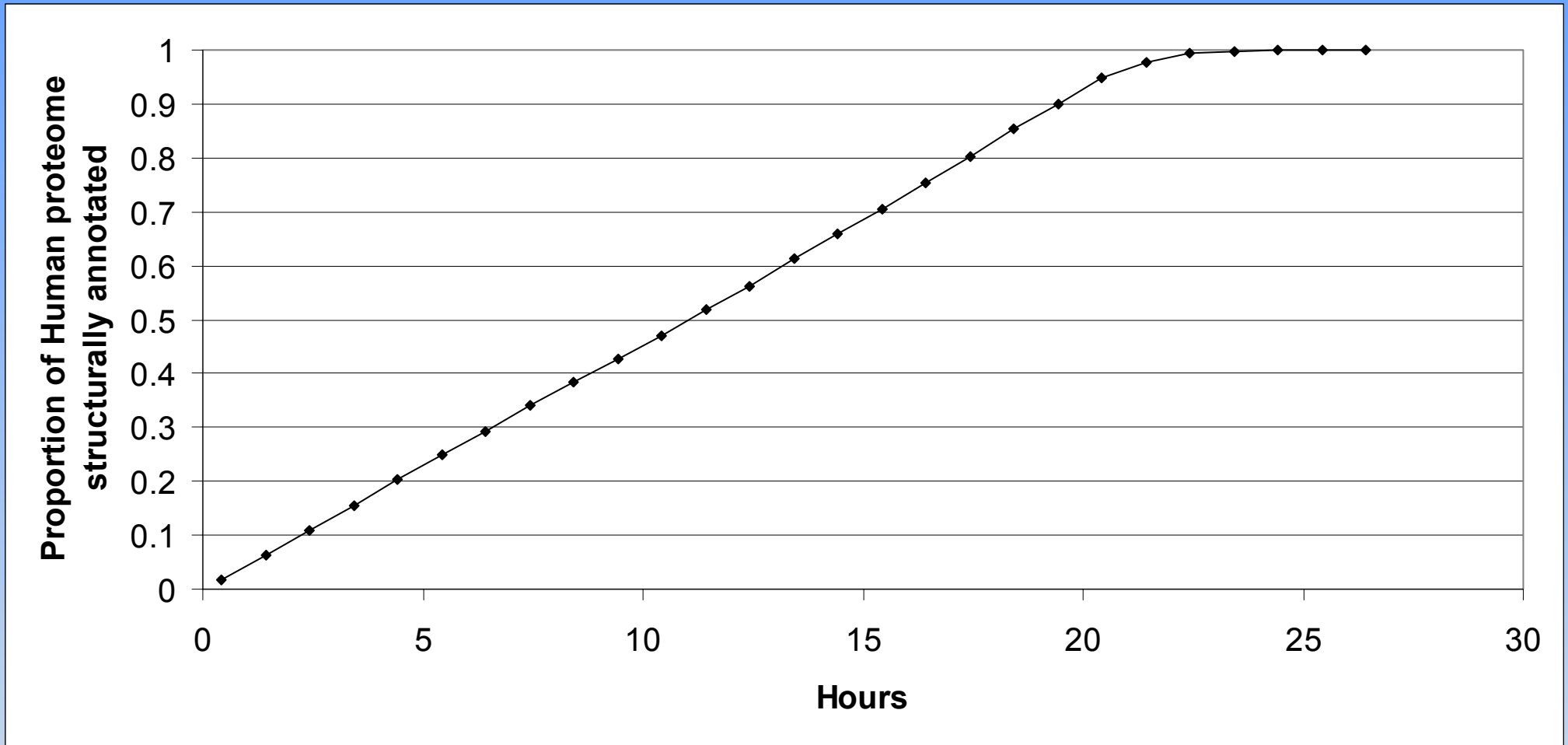


# Throughput



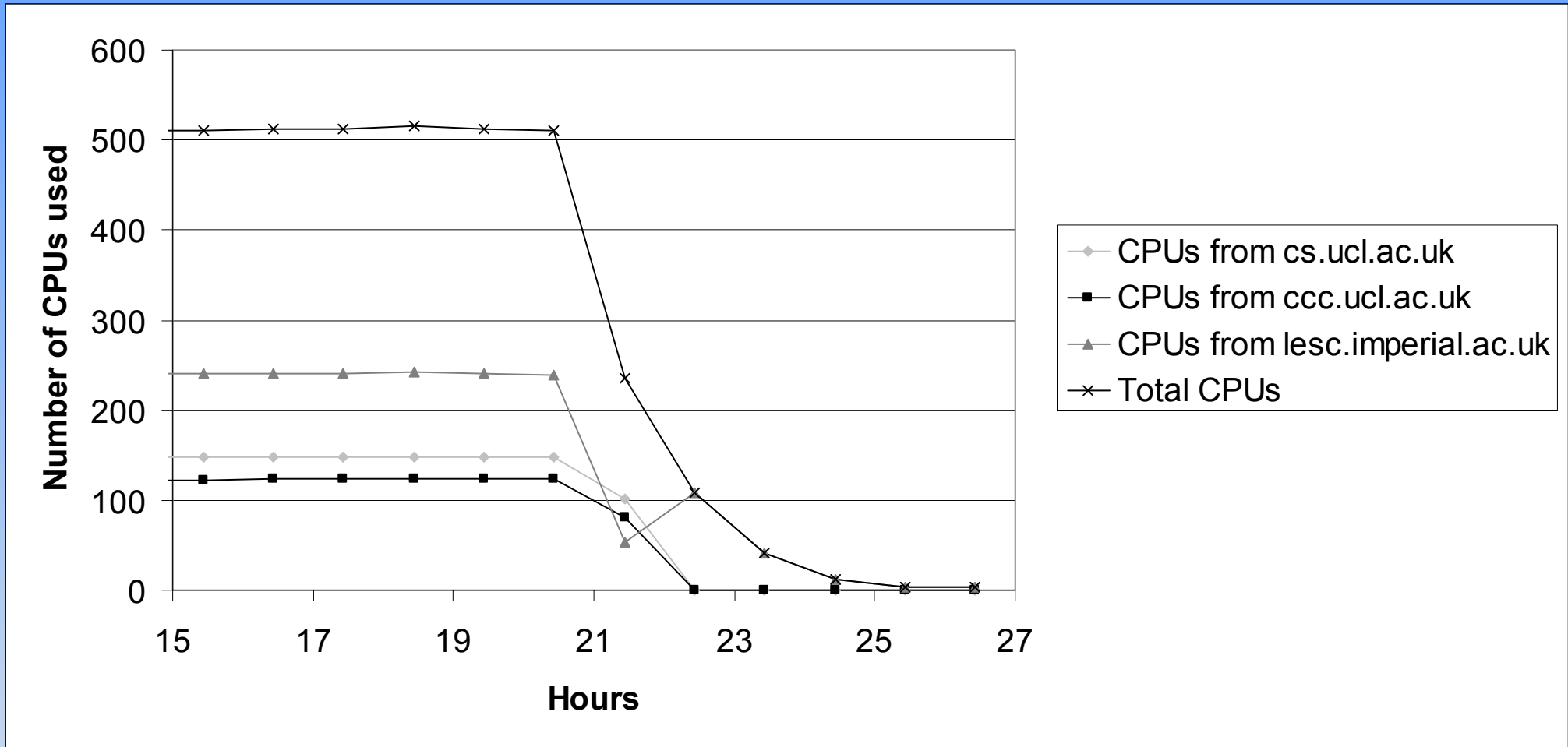
- Average: 1487 sequences per hour
- Max: 1617 sequences

# Throughput



- Linear prior to tailing off phase
- 99.9% in under 24 hours
- Total time 26 hours

# CPU usage



- Average: 504 CPUs per hour
- Max: 515 CPUs
- Tailing-off phase (CPUs > sequences left)

**ENSG00000170500 assigned to 2ane (an ATP-dependent protease) with with  $p < 0.001$ .**

Prior to the week that the fold library was created for the experiment, no structural assignment could have been made to this sequence region with that level of confidence.

According to a recent ENSEMBL search, a number of GO terms (GO:0006510 - ATP-dependent proteolysis and GO:0004176 - ATP-dependent peptidase activity) have been mapped to this entry via UniProt/RefSeq, which agree with this structural assignment.



# Acknowledgements

- Richard T Smith
- Kevin Bryson
- Søren-Aksel Sørensen
- David T Jones
  
- Keith Sephton at LESC
- William Hay at UCL
  
- Biotechnology and Biological Sciences Research Council
- Department of Trade and Industry (LJM, RTS)
- BioSapiens Network of Excellence (European Commission FP6 Programme contract number LSHG-CT-2003-503265).