

e-Protein: Vision, Challenges and Successes

www.e-protein.org



e-Protein



- A distributed pipeline for proteome annotation using grid technology



e-Protein



- Imperial College London
 - Prof M Sternberg (Molecular Biosciences)
 - Prof J Darlington (+ Dr Stephen Newhouse) (Computing)
 - RAs - Keiran Fleming,
 - Angela O'Brien, Murtaza Gulamali, Shikta Das
- University College London
 - Prof D Jones & Dr S Sorenson (Computer Science)
 - Prof C Orengo (Biochemistry)
 - RAs - Liam McGuffin, Stephano Street, Richard Smith
- European Bioinformatics Institute
 - Prof J Thornton, Dr E Birney, Dr A Robinson
 - RAs - Andreas Kahari, Tim Massingham



GRID

- Electricity GRID
 - Plug in your laptop, you get power but you do not have to consider where the source of the power is.
- Computer GRID
 - Submit an application, you get computing resources but you do not have to consider where the source of the compute power is.
- Early adopters
 - Physics community
 - Standard applications (particle physics) requiring huge compute resources and major data transfers **via established programs**.



E-Science

- Development of novel methods to exploit developments in computing and information technology to advance science.
 - Better
 - Faster
 - Cheaper
 - Easier
- GRID-based e-Science
 - e-Science involving disparate (and often remote) scientific activities.



BBSRC/DTI e-Science Pilot Projects 2001

- **BioSimGrid** (Sansom, Oxford)
 - A GRID database for biomolecular simulations
- **e-HTPX** (Nave, CCLRC)
 - An e-Science resource for high throughput structural biology
- **BDWorld** (Bisby, Reading)
 - problem-solving environment for global biodiversity: prototype and demonstrator
- **BASIS** (Kirkwood, Newcastle) *MRC co-funded*
 - Biology of Ageing: e-Science integration and simulation system
- **e-Protein**
 - A distributed pipeline for structure-based proteome annotation using GRID technology

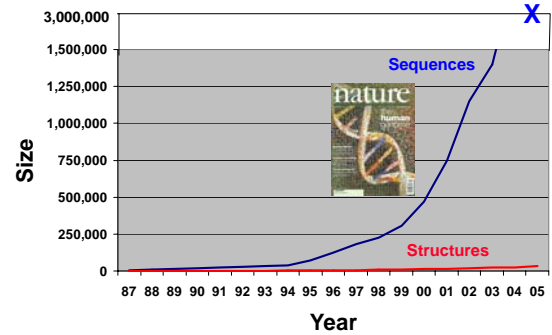


EPSRC 'Bio' E-science

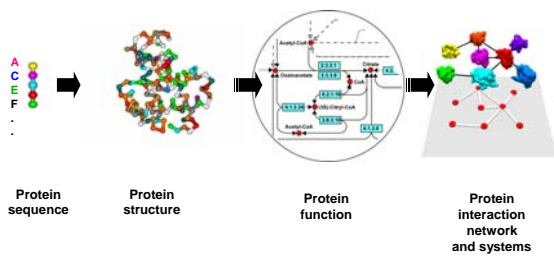
- MyGRID
 - Workflows for bioinformatics processes
 - Carole Goble et al
- DiscoveryNet
 - Workflows for high throughput 'data analysis
 - Yike Guo et al



Protein sequences and structures



Proteome Annotation

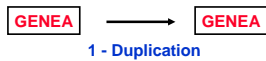


Gene Duplication

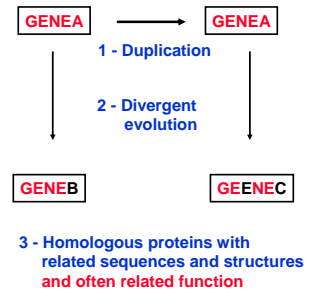
GENEA

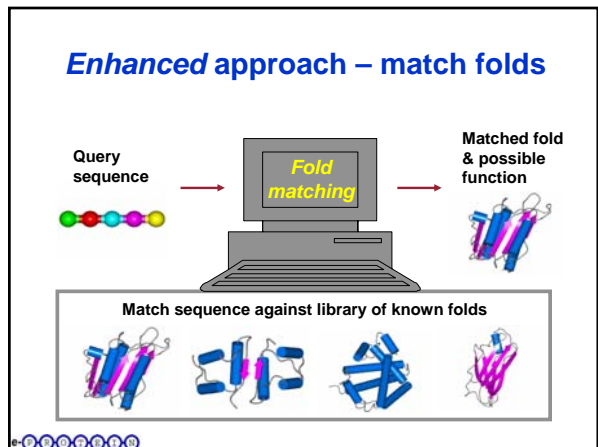
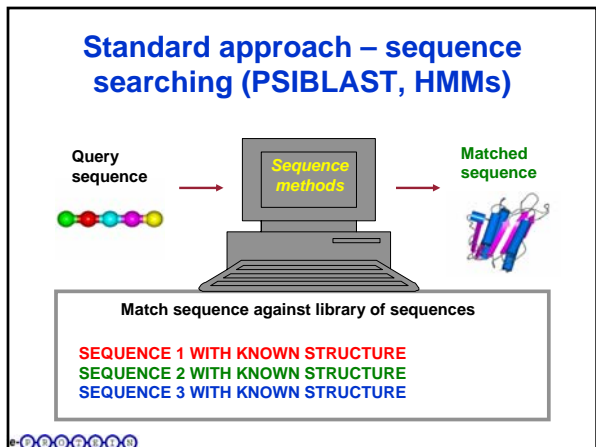
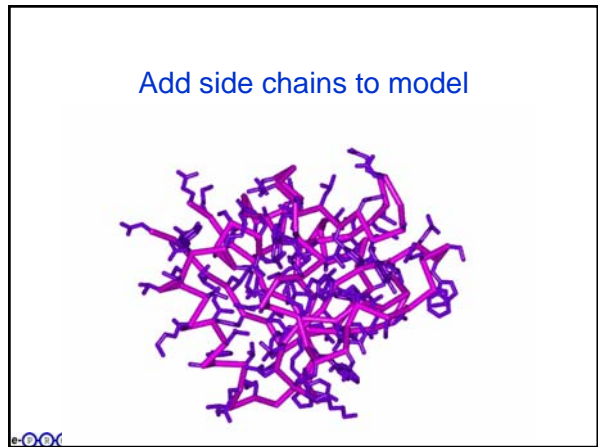
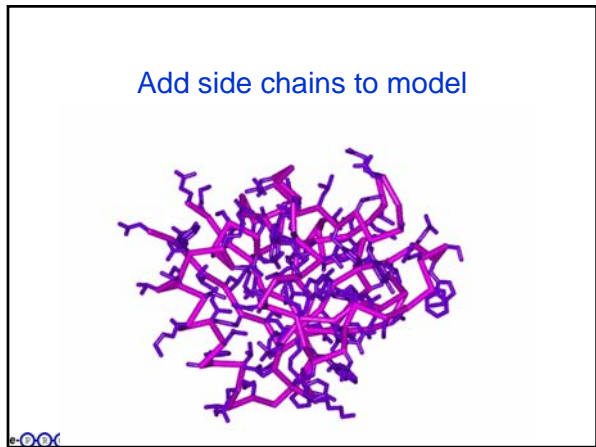
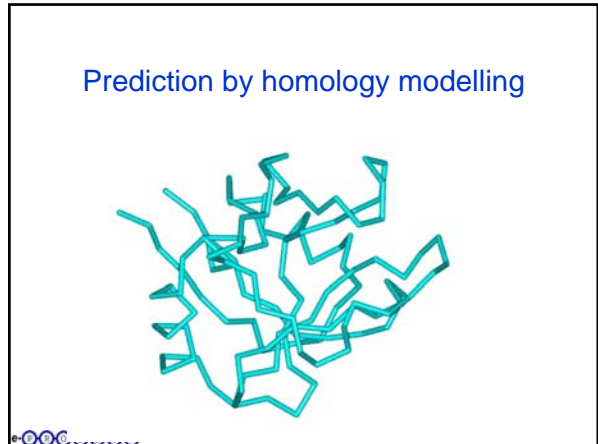
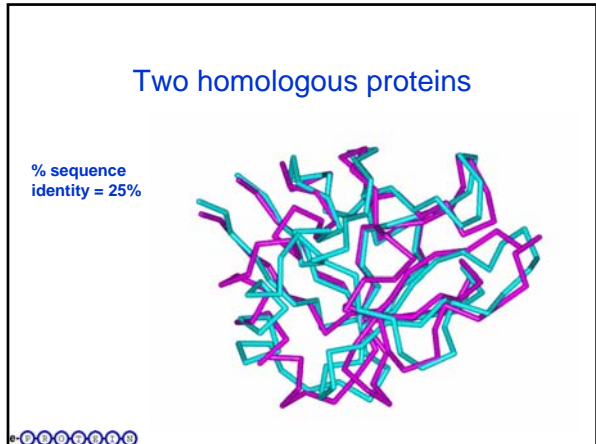


Gene Duplication

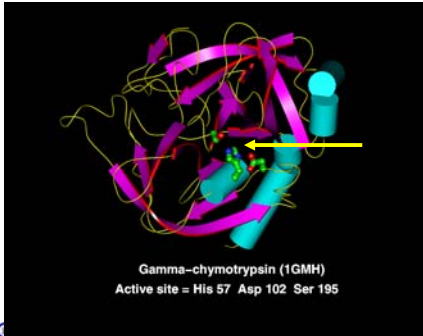


Gene Duplication





Map Functional Residues onto Structure



Function prediction aided by alignment

Known functional residues HKPSHAWRTKLYR

Function prediction aided by alignment

Known functional residues HKPSHAWRTKLYR
Match homologue HRPTHGRTLKYRT

Function prediction aided by structure

Known functional residues HKPSHAWRTKLYR
Match homologue HRPTHGRTLKYRT

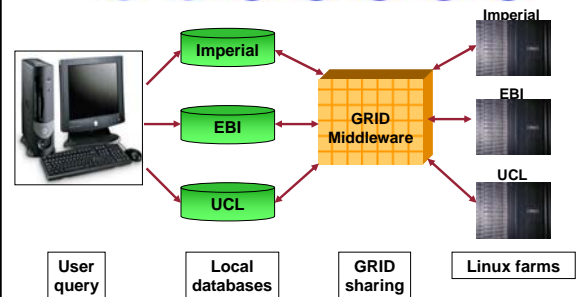
Do functional residues cluster?



e-Protein: Aims

- **Bioinformatics:** - To undertake a comprehensive structure-based annotation of the folds and function of proteins in model genomes at the three centres.
- **E-Science:** - To use Grid technology to provide the required computational power and e-science to enable the integration of the multi-site annotations.

e-PROTEIN



Main Computational resources

	Laboratory	Accessible Central Resources
Imperial (CS + Biol. Sci.)	50 P3 CPU	900 P4 / AMD CPU
UCL (CS + Biochem.)	700 P3 CPU 32 AMD CPU 16 SPARC CPU	- 500 P3 CPU 32 P4 (Myrinet)
EBI	-	200 P3

e-P-R-O-T-E-I-N

Linking compute power

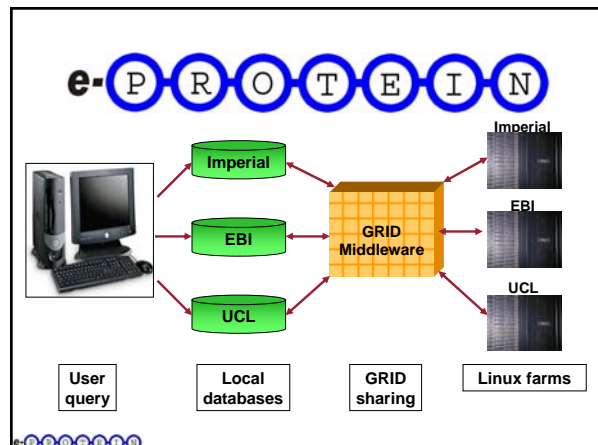
- **Challenge**
 - Available middleware (particularly GLOBUS) rapidly evolving but we required an advanced stable solution.
- **Solution**
 - Parallel developments by UCL and Imperial teams exploiting local expertise and several other GRID projects
 - GLOBUS not adopted and alternative solutions developed that are less monolithic.
 - UCL developed JYDE (Job Yield Distributed Environment)
 - Imperial developed ICENI evolved into GridSAM (OMII)

e-P-R-O-T-E-I-N

Successes of Distributed Computing

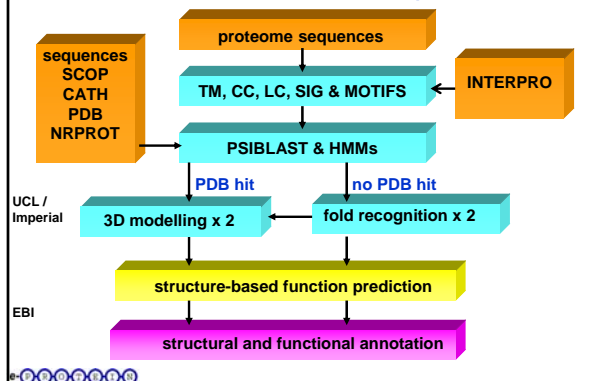
- **Annotation of the human proteome**
 - UCL used JYDE to access 500 processors in two different departments at UCL (Computing and Central Services) and at Imperial (LeSC) to annotate the 32,000 unique sequences in the human proteome in 24 hours.
- **Production Environment of human proteome annotation**
 - Imperial used GridSAM for 3D-Genomics annotation of 37,000 sequences in 44 hours using 800 processors at Imperial.
- **Use of non e-Protein resources**
 - Imperial have used the National GRID Service showing that the technology does not have to be carefully crafted to work.

e-P-R-O-T-E-I-N



e-P-R-O-T-E-I-N

e-Protein Annotation Pipeline

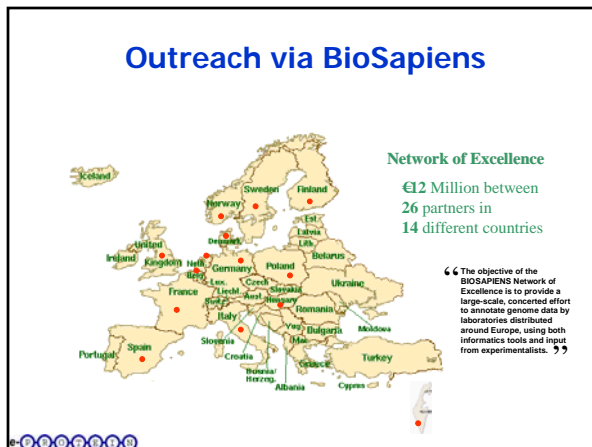
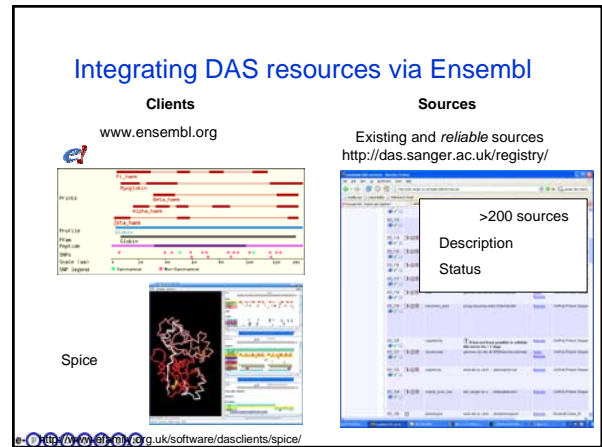
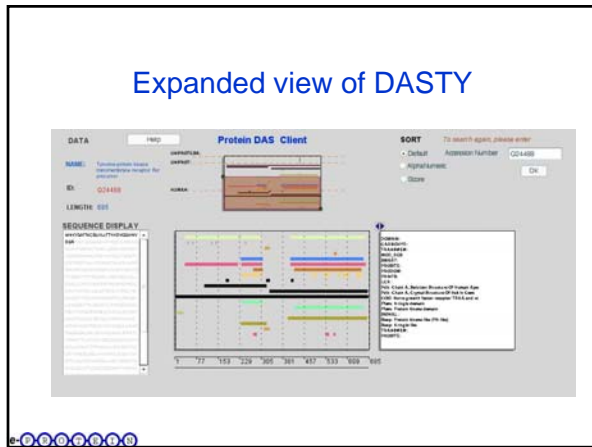
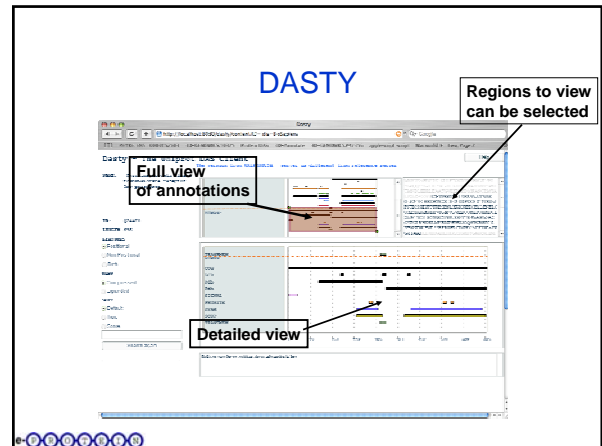
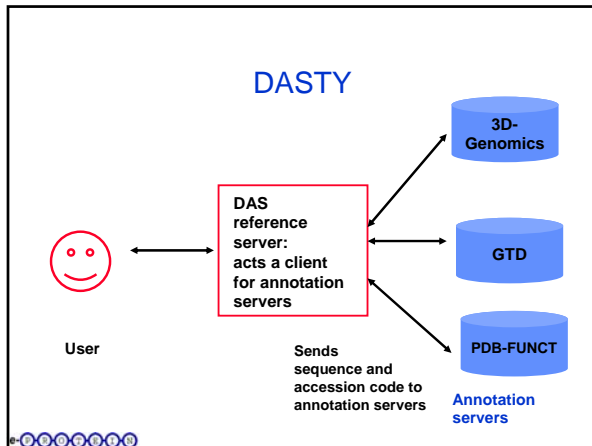


e-P-R-O-T-E-I-N

Integrating Annotation Databases

- **Challenge**
 - Each database has a complex and changing schema and it is not viable to integrate these schemas.
 - Individual groups wish to maintain the identity and character of their database.
- **Solution**
 - DAS (Distributed Annotation System)
 - Just agree how to refer and number each protein.

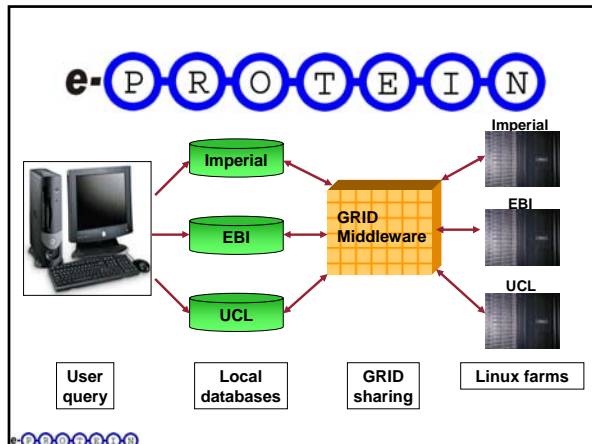
e-P-R-O-T-E-I-N



BioSapiens Genome Annotation

DNA Annotation	Proteome Annotation	Functional Annotation
<ul style="list-style-type: none"> •Gene definition/ alternative splicing •Regulators and promoters •Expression •Variation (haplotypes and SNPs) 	<ul style="list-style-type: none"> •Protein families, orthologues •Membrane proteins and ligands •3D protein structure •Post translational modification and localisation 	<ul style="list-style-type: none"> •Sequence and structure to function •Protein-protein complexes •Pathways and networks

Aim: New methods and new annotations
integrated via DAS



e-Protein Highlights: Bioinformatics

- Have used Grid technology to make available to academics the most comprehensive set of structural annotations and 3-D models across available genomes
- Grid technology will enable us to integrate structural and functional annotations
- Integrating resources has allowed far greater coverage than achievable from individual methods
- Will be able to provide new estimates of reliability that were not previously possible
- Will be able to deliver up to date annotations for key proteomes in realistic timescales over realistic resources

e-Protein Highlights: Grid Technology

- Have demonstrated and deployed Grid technology on a large-scale across several sites
- System developed to "production" level of stability
- To achieve this we have developed new resource management systems: JYDE and GridSAM
- Human proteome can be annotated in days using GRID technology
- Have refined and tested protein DAS protocols (e.g. consistency) to allow effective integration of protein annotation resources

Conclusion: Let's work together

- **Bioinformaticians:** e-science with a GRID-based approaches provides powerful and viable methodologies for enhanced bioinformatics in terms of access to computing resources and data integration.
- **Computer scientists:** bioinformatics provides a challenging and worthwhile application domain that will stimulate the development of novel and useful computational solutions.



e-Protein



- **Imperial College London**
 - Prof M Sternberg (Molecular Biosciences)
 - Prof J Darlington (+ Dr Stephen Newhouse) (Computing)
 - PDRA - Keiran Fleming,
 - Angela O'Brien, Murtaza Gulamali, Shikta Das
- **University College London**
 - Prof D Jones & Dr S Sorenson (Computer Science)
 - Prof C Orengo (Biochemistry)
 - PDRA - Liam McGuffin, Sterphano Street, Richard Smith
- **European Bioinformatics Institute**
 - Prof J Thornton, Dr E Birney, Dr A Robinson
 - PDRA - Andreas Kahari, Tim Massingham