Young Investigators Symposium
Oak Ridge, Tenessee Oct 2008

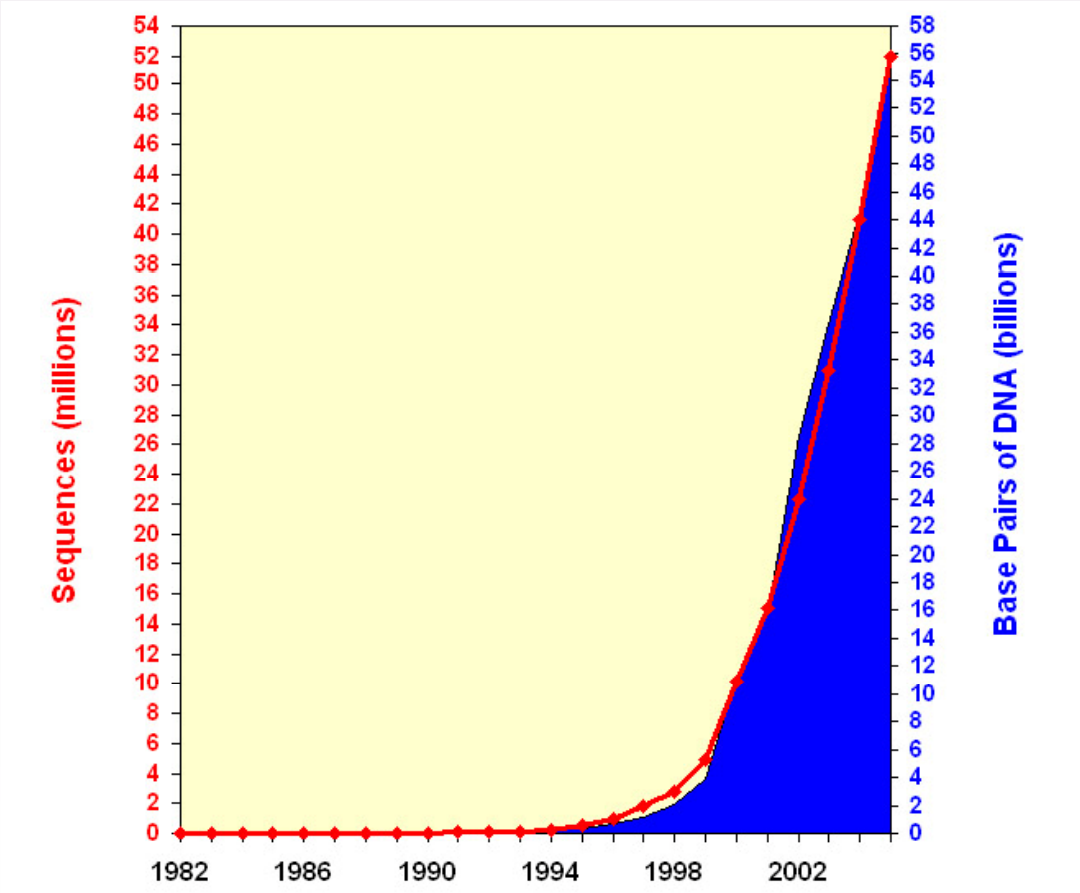
Bioinformatics Challenges:
Predicting Functions for the human proteome

Anna Lobley PhD Student

Post Genomics and Systems Complexities

In basepairs
56 billion

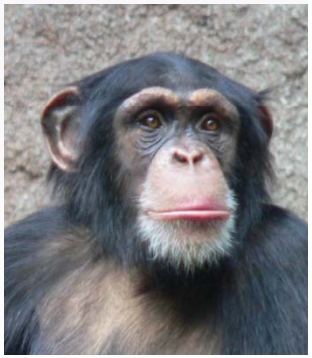
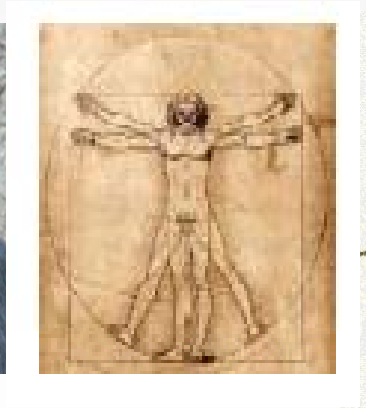
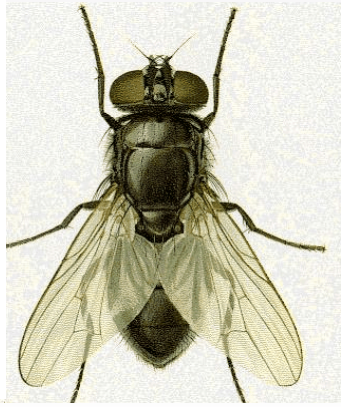
In sequences
52 million



Impact of High Throughput Sequencing on Biological Database Growth

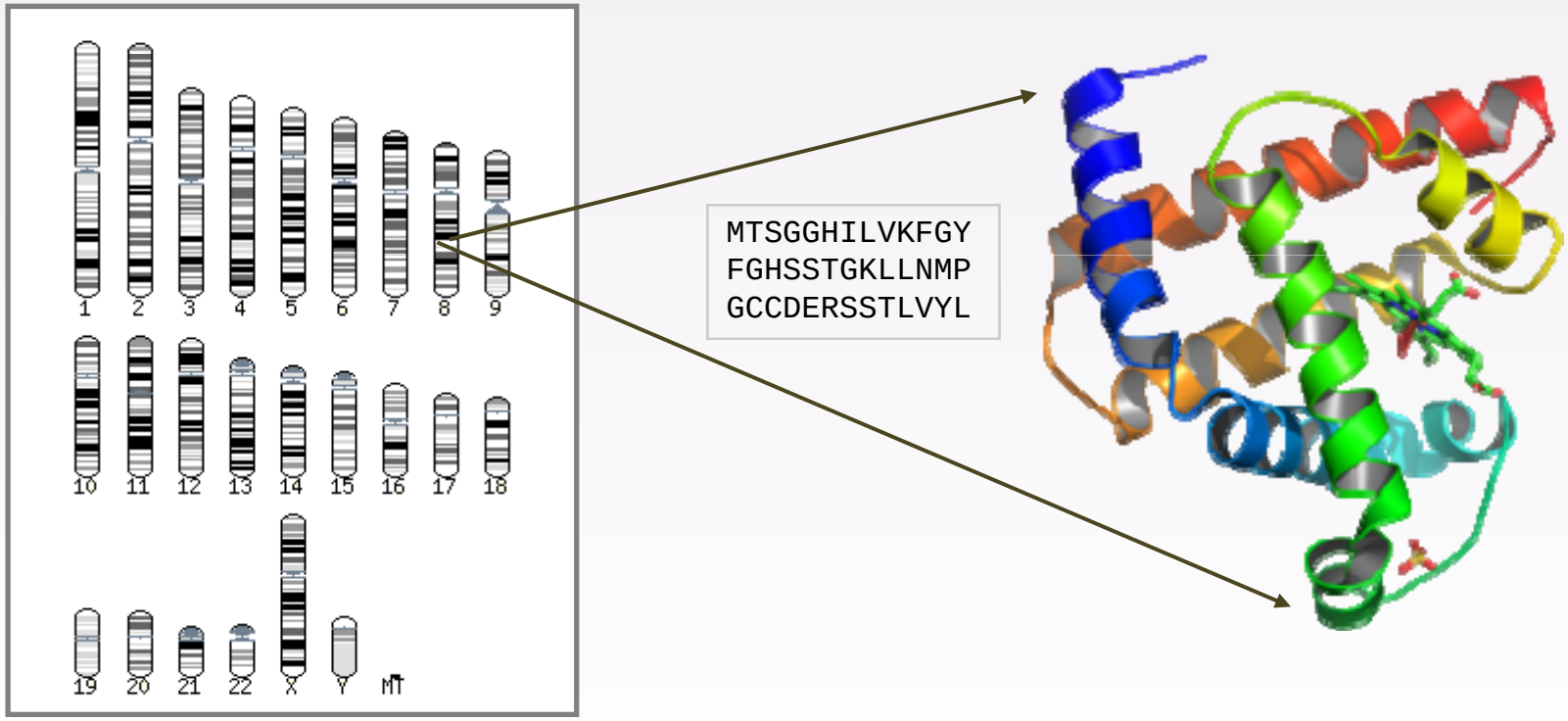
Components of Genomes Known
Now can model behaviour

There are currently over 800 fully sequenced genomes

Attention turns to the proteins – the work horse components of living cells

Genomes to Proteomes



3,253,037,807 bp

25-30K proteins in the human proteome

Only 3400 of structures are available for human proteins

Making use of protein sequence information:

- Protein identification and function
- Protein - protein interactions
- Model whole pathways
- Whole system behaviour

Protein Identification and Function

Protein Function

- What does the protein do
- Where does it perform its function(s)
- How does it perform its function(s)

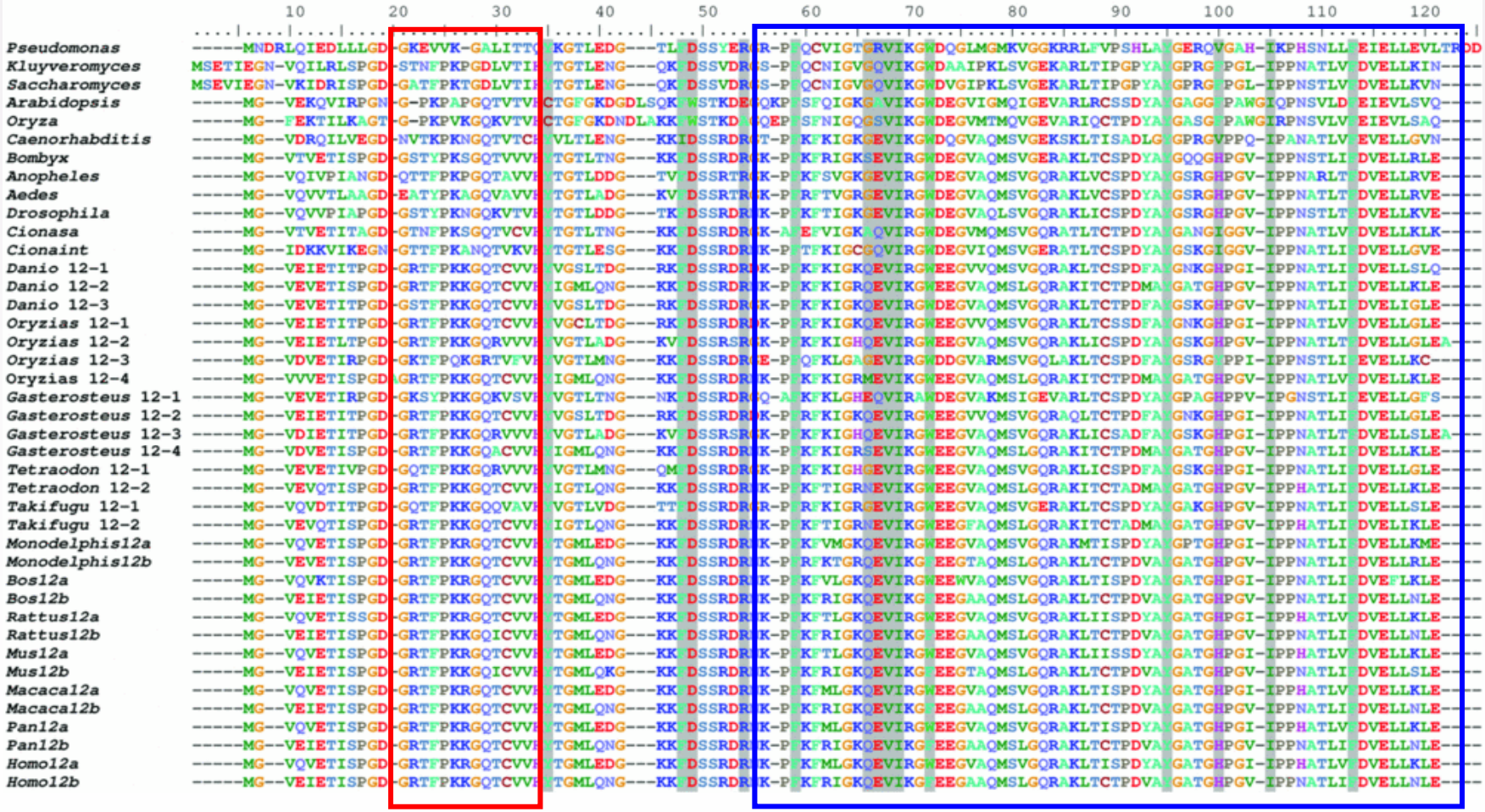
Computational Strategy

One sequence alone is not useful

- Evolutionary history very useful
- Neighborhood of sequence is useful
- Create sequence profile models
- **P**osition **s**pecific **i**terated (Psi-BLAST)

Basic **L**ocal **A**lignment **S**earch **T**ool

Extracting signals from noise



localisation motif

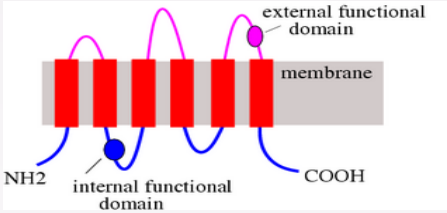
structural domain

Function Prediction Approach

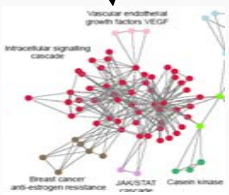
Query

Psi-BLAST model

Query	MTFDDWFGHYKI
Mouse	MHKKLTDFSTREE
Rat	- - LDVTQDTK - -
Fly	MDPGFTSG – K - -
Zebrafish	M I LGDTRSSYW -
Worm	- - KYNTDVC- - -

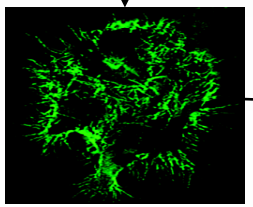


transmembrane



Protein interactions

structure

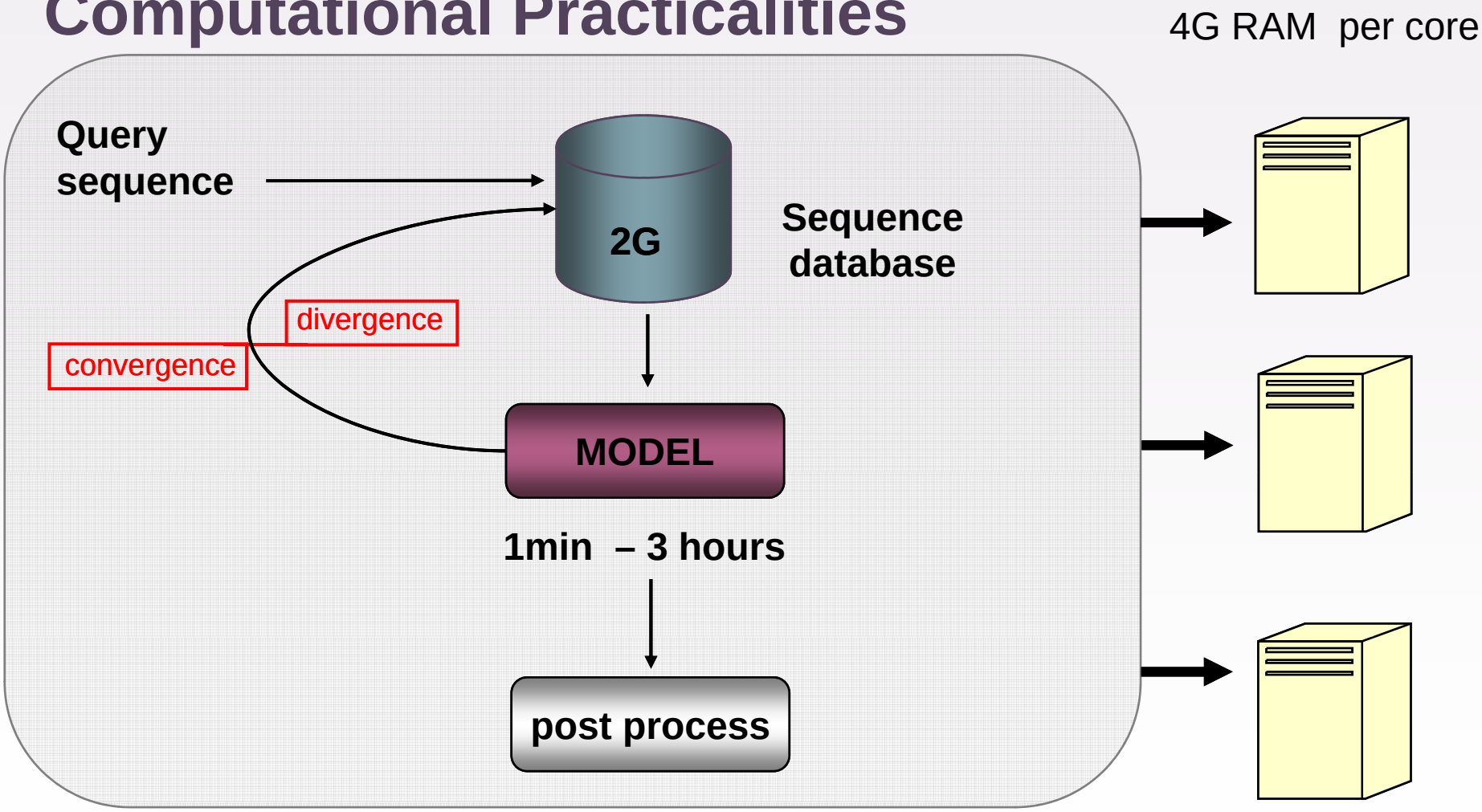


localisation

Support vector machine

DNA binding

Computational Practicalities



Embarassingly parallel application: many serial jobs

Function Prediction for the human proteome

PSI-BLAST models for 30 000 proteins

- 3 months processing on 200 cores

Function prediction

- Grid searches to optimise SVM parameters
  Ensemble classifiers (5 per function) rbf kernel
- 3 months optimising parameters for Support Vector machines to classify over 300 functions
- 322 feature space with 14k examples

Lobley, A.E., Nugent, T., Orengo, C.A. & Jones, D.T. (2008) FFPred: an integrated feature-based function prediction server for vertebrate proteomes. Nucleic Acids Res.

Lobley, A., Swindells, M.B., Orengo, C.A. & Jones, D.T. (2007) Inferring function using patterns of native disorder in proteins. PLoS Comput. Biol. 3:e162.

Current work

- Predict structure, function, interactions of all
  5.5 million proteins (entire protein universe)

- >1 million processed to date:
  180-200K sequences per day on Legion cluster
  target time 5 weeks

Key Challenges

PSI-BLAST models must be kept up-to date

- 1 new sequence results in very different profile
- database grown ½ million in last 6 months

Work in protein pair space

- 450 million matrix elements for human

Machine learning on large datasets

- ~ 1million datapoints with complex kernels and higher dimensional space

Acknowledgements

Legion and is-rec compute teams

Biosapiens network of excellence

BBSRC Case Studentship (BioFocus DPI)