Computational biology exploration of the enzymatic diversity of an uncharacterised prokaryotic protein family

Adam Alexander T. SMITH
Marcel SALANOUBAT
Jean WEISSENBACH
Claudine MEDIGUE
David VALLENET

UMR 8030
Genoscope, Institut de Génomique, CEA

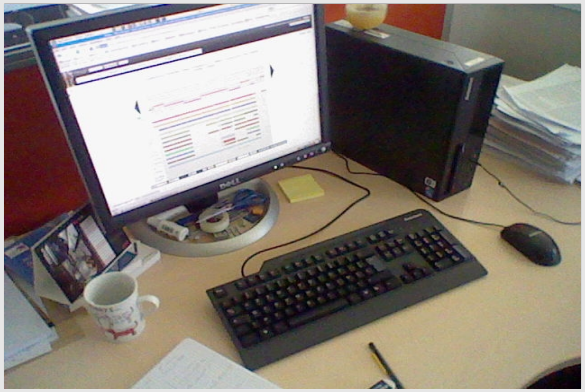
September 2010

Discovery of a gene for an orphan activity

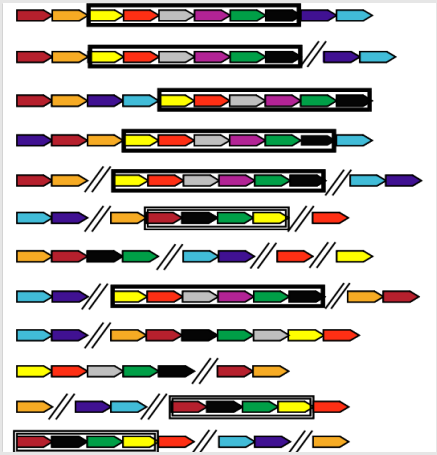
Approx. 30% of EC numbers defined by the IUBMB are still sequence-orphans.

The Genoscope has started a transversal initiative to identify genes for these activities.

1 pilot project



Bioinformatics analysis of an incomplete metabolic pathway led to ...



...the discovery of a conserved genomic context, hinting a possible activity for an uncharacterised protein family...

biochemical assays

What about the others? (approx. 660)

... which was then experimentally confirmed on **some** (12) family members.

Overview of the BKACE Pfam family

DATA :
Family of proteins defined by the presence of a conserved Pfam domain.
Illustrated : alignment of DOMAIN over 663 proteins

OBJECTIVE :
explore metabolic diversity of all proteins

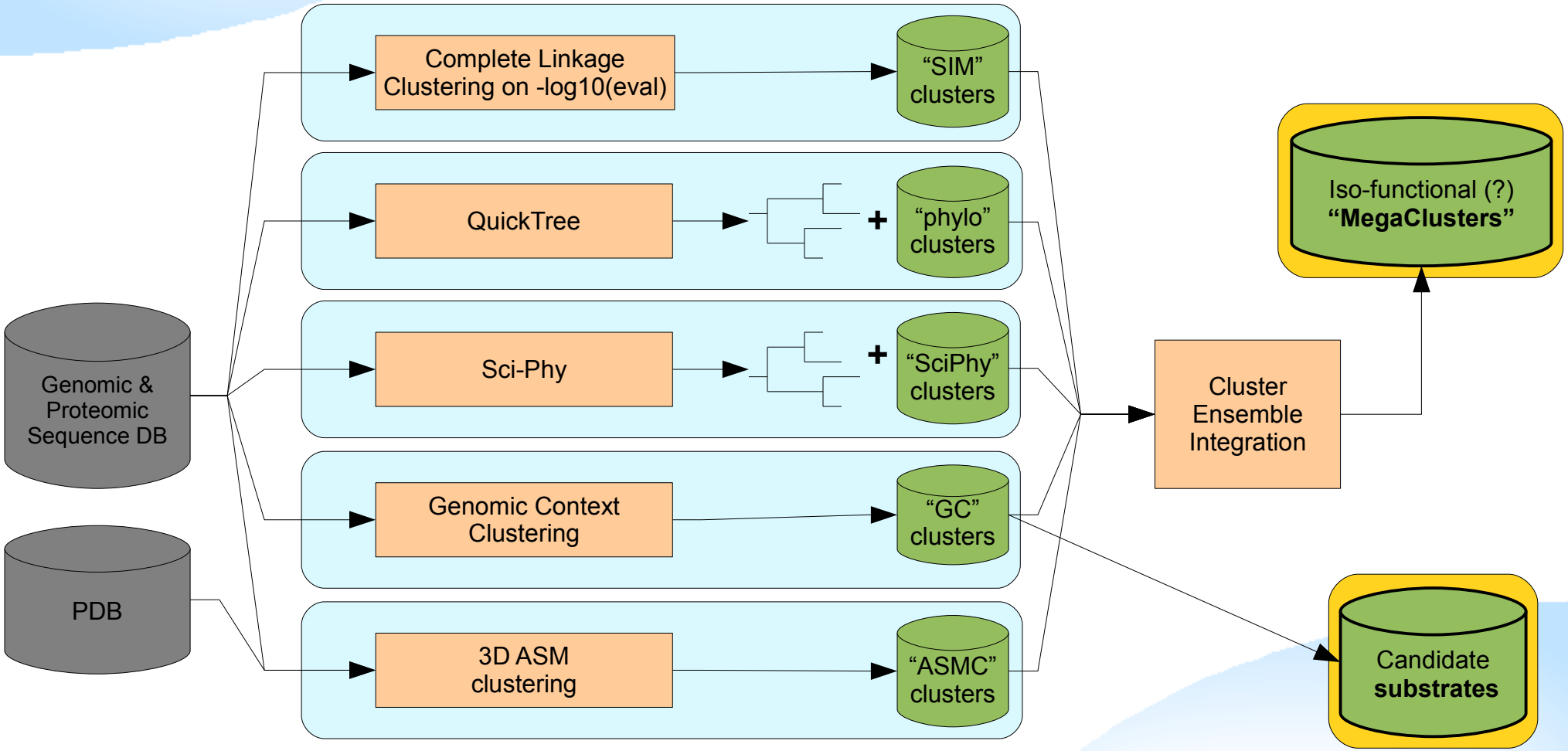
GENERAL STRATEGY :

- using computational biology :
  - cluster proteins into **iso-functional** groups
  - use diverse functional homology clues
  - propose **candidate substrates** for each cluster

- biochemical assays :
  - select representatives for each cluster
  - test known & proposed activities

Pfam : http://pfam.sanger.ac.uk/
"The Pfam protein families database" R.D. Finn & al., Nuc. Acids Res. (2010) Database Issue 38:D211-222

Clustering strategies applied to the BKACE family



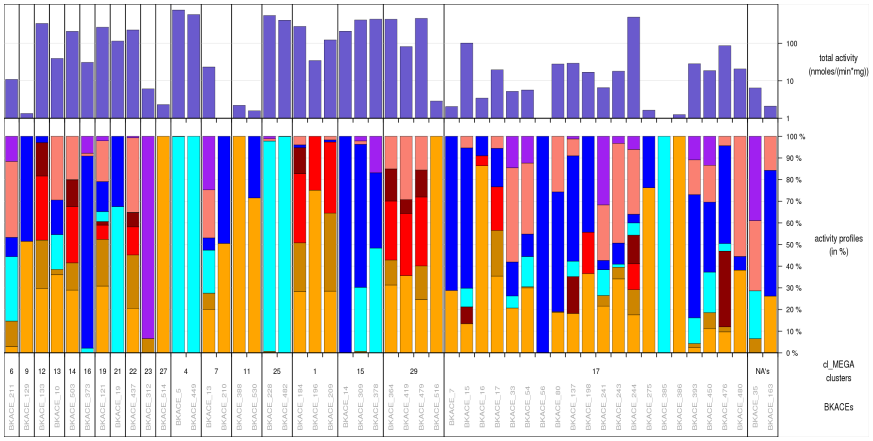
Clustering & Biochemical results

Proposed candidate substrates :

- 4 new candidates proposed using metabolic contexts
- 7 new candidates proposed using substrate similarity
  - ➔ 8 found as potential substrates so far

MegaClustering :

- 88 MegaClusters
- heterogeneous cluster sizes (3-130); no singletons
- 1 “dustbin” cluster



Biochemistry :

- 130 BKACE attempted
- 62 BKACE cloned and tested for proposed activities
- 52 positive hits
- too few cloning successes (high GC content) per MegaCluster
  - ➔ Impossible to validate MegaClustering strategy
  - ➔ New candidates currently being tested

Thankfully :

- ➔ coherence of various clusterings i.r.t. biochemistry
- ➔ explored a varied part of family's functional space

Perspectives

- Finish new clones & biochemical assays

- Assess coherence of different clusterings (GC, ASMC, MegaClustering)

- Assess usefulness of metabolic context pooling for substrate proposition

- Establish generalised strategy / protocol

- Try out on new families

Many thanks to ...

Laboratoire d'Analyses Bioinformatiques pour la Génomique et le Métabolisme

Claudine MEDIGUE

David VALLENET

Alexandra CALTEAU

Laboratoire de Chimie Organique et Biocatalyse

Anne ZAPARUCHA

Carine VERGNE

Thésaurus Métabolique

Marcel SALANOUBAT

Véronique DE BERARDNIS

Annett KREIMEYER

Alain PERRET

Laboratoire d'Analyses Bioinformatiques des Séquences

François ARTIGUENAVE

Racquel CARDOSO DE MELO MINARDI

Karine BASTARD

Hungry for more details ?

Come discuss the strategy in front of our poster (#23)

Thank you for your attention !