# The CAZy database, a tool for enzyme discovery

**Bernard Henrissat**

Lab: Architecture et Fonction des Macromolécules Biologiques

CNRS and Aix-Marseille University, Marseille, France



# Carbohydrates

- **Not directly encoded by DNA**
- interesting **when attached** to each other (glycans)
- amazing **stereochemical diversity** despite similar/boring composition
- hugely **abundant** (photosynthesis), source of carbon for practically all living organisms
- large **applied interest :** wood, pulp & paper, agriculture, food, feed, drinks, phytopathogens & biocontrol, biofuels & green chemistry, biotechnology, health & medicine, (mal)nutrition, etc

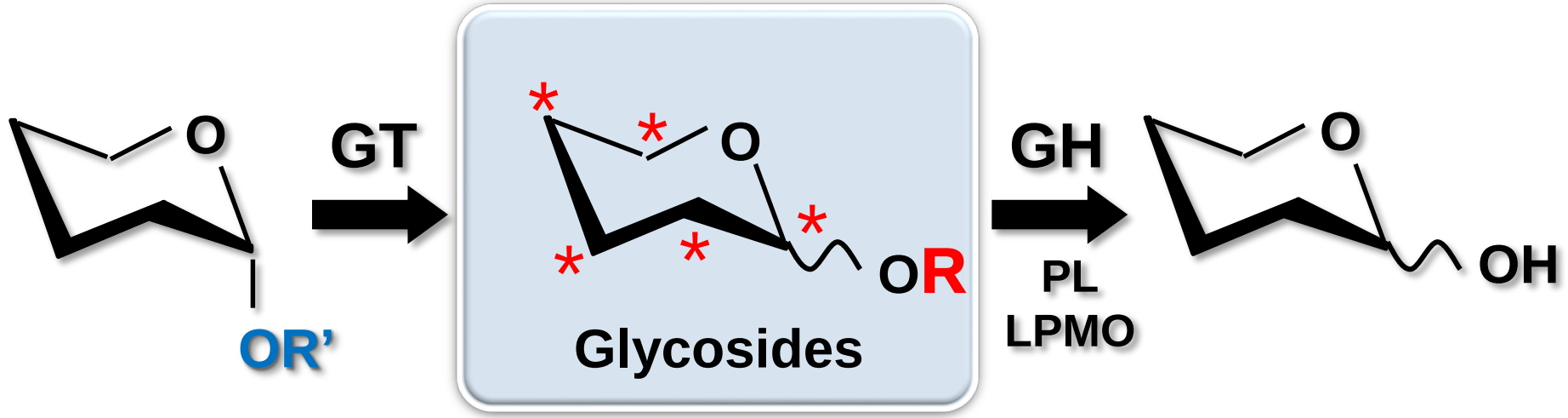
# For many people the only exciting forms of carbohydrates are food related








# My work is to explore the link between carbohydrates and CAZyme sequences




*Breaking a sugar code : to realize the potential offered by genomics, we need to establish ways to accurately predict specificity of carbohydrate-active enzymes from their amino acid sequences*

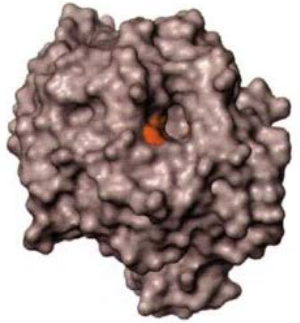
# Carbohydrates (continued)

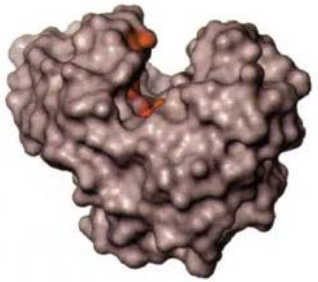
- Virtually **any molecule of life** can be glycosylated at some stage (lipids, nucleic acids, antibiotics, steroids & hormones, proteins ... and of course sugars themselves)
- Many ways to link sugars together : there is an **astronomical number** of oligo- and polysaccharide structures in Nature
- Consequence : there is an **enormous diversity** of CAZymes

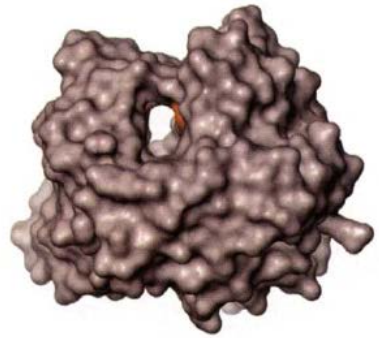
- The stereochemical features of carbohydrates enable proteins to act upon them **selectively** → immense variety of biological functions

# Features of carbohydrate-active enzymes

- To make use of the **amazing structural variety of carbohydrates**, Nature was able to evolve proteins able to **selectively** assemble and breakdown glycoconjugates, oligo- and polysaccharides
- Selective recognition is achieved by active sites offering a **large complementary surface** (pocket, cleft, tunnel) which confers the desired specificity
- Many more carbohydrate structures than there are protein folds : acquisition of different specificities on a limited number of ancestral scaffolds has left **traces** in the aminoacid sequence of CAZymes







# Principle : compare amino acid sequences and group enzymes in families of related sequences :

- 1991 – now:  a classification of glycosidases
- 1997 – now: glycosyltransferases
- 1999 – now: carbohydrate esterases
- 1999 – now: carbohydrate-binding modules
- 2010 – now: polysaccharide lyases
- 2013 – now: auxiliary activities (redox enzymes)

# CAZy : the Carbohydrate-active enzymes database (www.cazy.org)

- **Families** of enzymes and protein domains that assemble, cut and bind **complex carbohydrates**
- Launched **Sept 1998** - Updates every 3-4 weeks

**→ CAZy will be 20 years old this year !**

- Underlying classification work started in **1989-1991**
- GHs: 1991; GTs: 1997; CEs, PLs and CBMs: 1999; AA: 2013
- Total of **~400 families***
- Data source : NCBI **daily releases of GenBank** (amino acid sequences)

- Analysis using BLASTp & home made HMMs, **plus human curation**
- **Public side*** (>9,000 bacterial, ~300 archaeal, >200 eukaryotic genomes)
- **Private area** for collaborations on draft genomes (~1,500 bacterial, ~1,500 eukaryotic genomes including ~1,400 fungi)
- CAZyme families have an extremely **rich functional information** content
- CAZymes inform on the **lifestyle and carbon utilization pathways** of organisms

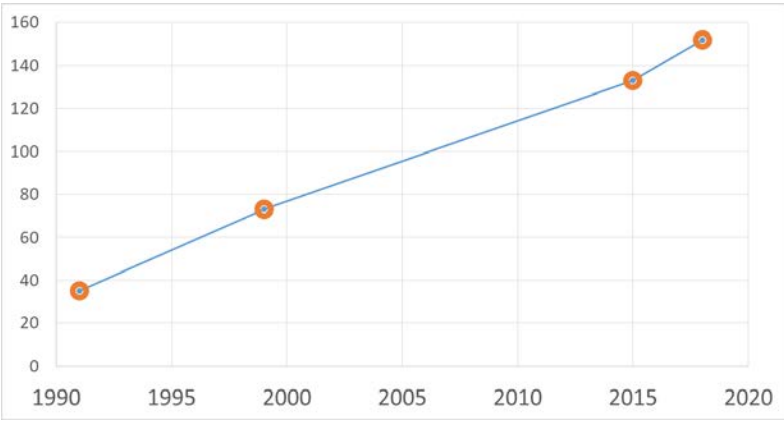
* February 2018

# CAZy : a knowledge base serving the Glycobiology community

- For CAZymes, activity means **substrate and/or product specificity**
- For each family CAZy gives (and updates !) the **known activities**, the nature of the **catalytic mechanism** and of the **catalytic machinery**
- **Statistics** on the family
- **Listings** for all organisms, or just Archaea or Bacteria or Eukaryota
- Listing of the enzymes with **known 3-D structures** and display of the **ligand information**
- Listing of those enzymes that have been **experimentally characterized**
- For some families : breakdown in **subfamilies**

- CAZy is strongly engaged in Glycobiology and Microbiology **research**
- CAZy is operated by a **small group of dedicated people** with deep knowledge of carbohydrates and their enzymes as **their research tool**
- CAZy is **not directly funded to perform any of its activities**
- **Community cooperates with CAZy** to alert on new activities, and to request new family numbers
- Complemented by **CAZypedia**, a community-driven encyclopedic resource on carbohydrate-active enzymes (www.cazypedia.org)

# Number of GH families




| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 11 | | 31 | | 51 | | 71 | 81 | 91 | 101 | 111 | 121 | 131 | 141 | 151 |
| 2 | 12 | 22 | 32 | 42 | 52 | 62 | 72 | 82 | 92 | 102 | 112 | 122 | 132 | 142 | 152 |
| 3 | 13 | 23 | 33 | 43 | 53 | 63 | 73 | 83 | 93 | 103 | 113 | 123 | 133 | 143 | |
| 4 | 14 | 24 | 34 | 44 | 54 | 64 | 74 | 84 | 94 | 104 | 114 | 124 | 134 | 144 | |
| 5 | 15 | 25 | 35 | 45 | 55 | 65 | 75 | 85 | 95 | 105 | 115 | 125 | 135 | 145 | |
| 6 | 16 | 26 | 36 | 46 | 56 | 66 | 76 | 86 | 96 | 106 | 116 | 126 | 136 | 146 | |
| 7 | 17 | 27 | 37 | 47 | 57 | 67 | 77 | 87 | 97 | 107 | 117 | 127 | 137 | 147 | |
| 8 | 18 | 28 | 38 | 48 | 58 | 68 | 78 | 88 | 98 | 108 | 118 | 128 | 138 | 148 | |
| 9 | 19 | 29 | 39 | 49 | 59 | | 79 | 89 | 99 | 109 | 119 | 129 | 139 | 149 | |
| 10 | 20 | 30 | | 50 | | 70 | 80 | 90 | 100 | 110 | 120 | 130 | 140 | 150 | |

1991 1999 2015 2018

*More than 4 times more families known in 2018 than in 1991: the share of carbohydrate-active enzymes in genomes is growing steadily !*

# The big challenge: functional predictions

- Many researchers utilize **CAZy family membership as a functional prediction**
- But most of our **families group together different EC numbers** (sometimes more than 20 … and all activities are not known)
- Capture of functional information:
  - Nowadays **biochemists no longer use EC numbers**
  - EC numbers : what are they ? What were they made for ? **Are they adapted to Bioinformatics ?**
- **Naming system(s)** for protein families and the **propagation** of the functional information by sequence similarity

# How are genes annotated by genome annotators ? How are functions predicted ?

- By comparing the sequences of the putative proteins to all proteins/profiles in a sequence or a profile database **at a given time**
- Essentially by inspection of the top **BLAST** hits or the top scoring family **HMM**
- Mostly by **different** people, using **different** criteria, **different** methods or **different** thresholds
- Whilst this is perhaps not a problem for a number of proteins, inspection of literature shows that there are **serious problems** with the annotation of CAZymes

# Features of the CAZy families

- Conserved molecular mechanism
- Conserved catalytic residues
- Conserved 3-D fold

} ***Predictive power***

- **Varying substrate / product specificity**
- **Variable modular structure**

→ ***Problems in functional prediction***

# Two *Volvariella volvacea* genomes, two teams, two methods, different results

**Chen et al. PLoS One. 2013; 8(3): e58780**

- Published: **March 12, 2013**
- CAZymes predicted with **CAT**
- « ranks 7th among 15 fungi with sequenced genomes »
- « the composition of glycoside hydrolases in *V. volcacea* is dramatically different from other basidiomycetes »

**Bao et al. PLoS One. 2013; 8(3): e58294**

- Published: **March 19, 2013**
- CAZymes predicted with **dbCAN**
- « ranks 3rd among 5 basidiomycetes »
- « the genome of *V. volvacea* has many genes that code for enzymes which are involved in the degradation of cellulose, hemicellulose, and pectin »

CAT : Park et al. (2010) Glycobiology 20:1574-1584
dbCAN : Yin et al. (2012) Nucleic Acids Res. 40:W445-451

# Comparison of the GH family profiles

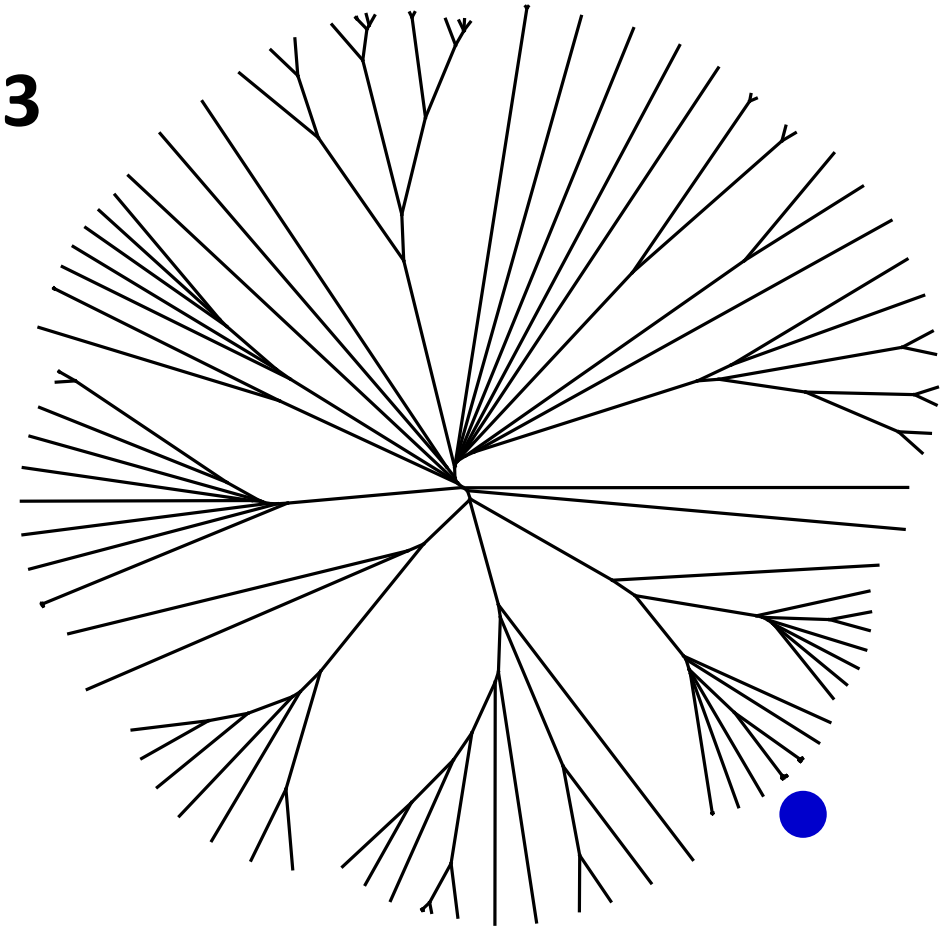
| | Chen | Bao | |
|---|---|---|---|
| GH1 | 3 | 3 | |
| GH2 | | 2 | |
| GH3 | 12 | 11 | |
| GH5 | 6 | 17 | contains cellulases |
| GH6 | 4 | | contains cellulases |
| GH7 | 11 | 14 | contains cellulases |
| GH9 | 1 | 1 | |
| GH10 | 20 | 19 | |
| GH12 | 2 | 2 | |
| GH13 | 9 | 7 | |
| GH15 | 3 | 5 | |
| GH16 | 2 | 21 | |
| GH17 | 1 | 3 | |
| GH18 | 12 | 11 | |
| GH20 | 1 | 1 | |
| GH23 | 1 | 1 | |
| GH24 | 1 | | |
| GH25 | 3 | | |
| GH27 | 2 | 1 | |
| GH28 | 3 | 3 | |
| GH30 | 2 | 2 | |
| GH31 | 6 | 6 | |
| GH35 | 4 | 4 | |

| | Chen | Bao |
|---|---|---|
| GH37 | 1 | 2 |
| GH38 | 1 | 1 |
| GH43 | 12 | 8 |
| GH47 | 7 | 7 |
| GH51 | 3 | 3 |
| GH53 | 1 | 1 |
| GH55 | | 5 |
| GH61 | 31 | 30 |
| GH63 | 1 | 1 |
| GH71 | 2 | 4 |
| GH72 | 1 | 1 |
| GH74 | | 1 |
| GH79 | | 5 |
| GH85 | 3 | |
| GH88 | 5 | 1 |
| GH92 | 4 | |
| GH95 | | 2 |
| GH105 | | 2 |
| GH109 | 10 | |
| GH115 | | 3 |
| GH125 | | 1 |
| GH128 | | 4 |
| Total | 191 | 216 |

Same result
Diff ≤ 2
Diff ≥ 3

# Problems with functional prediction

**Family GT83**

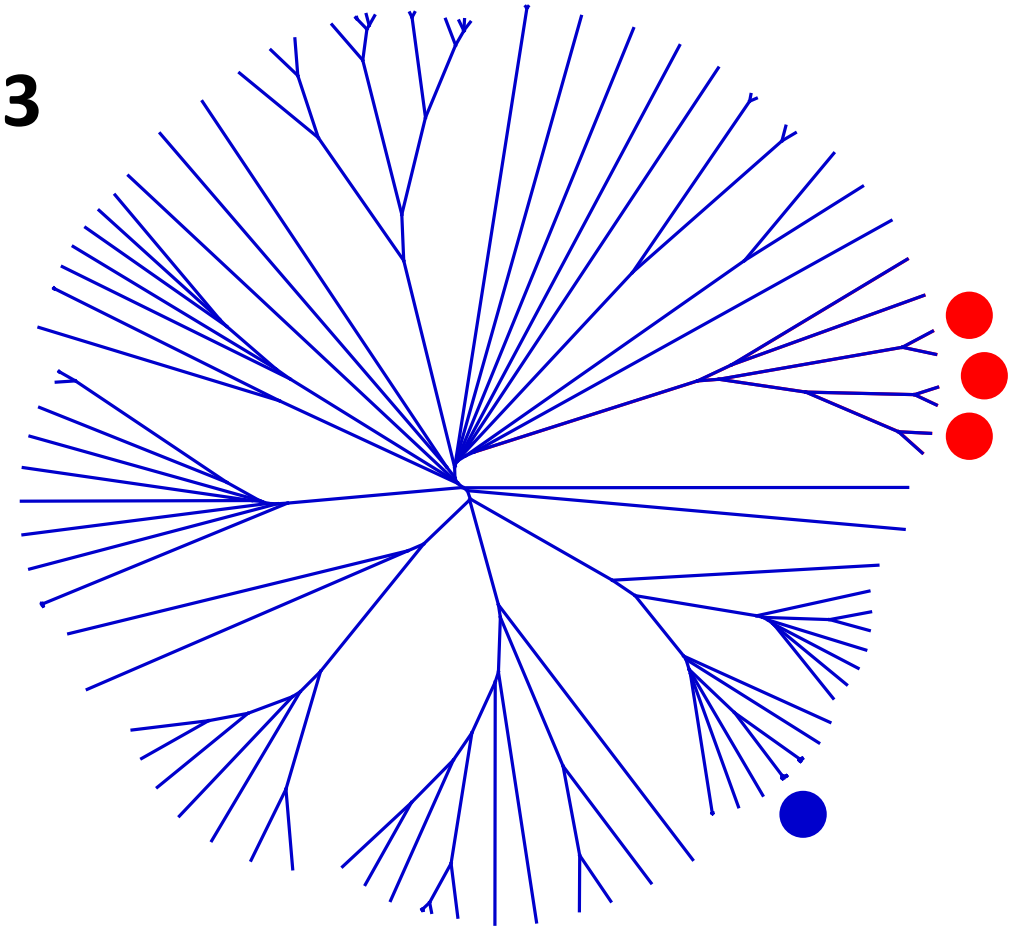


● **undecaprenyl phosphate-L-Ara4N:4-amino-4-deoxy-β-L-arabinosyltransferase (EC 2.4.2.43)**

- **C. Raetz, et al. (2006) :** dodecaprenyl phosphate- β-galacturonic acid: α-galacturonosyl transferases (EC 2.4.1.-)

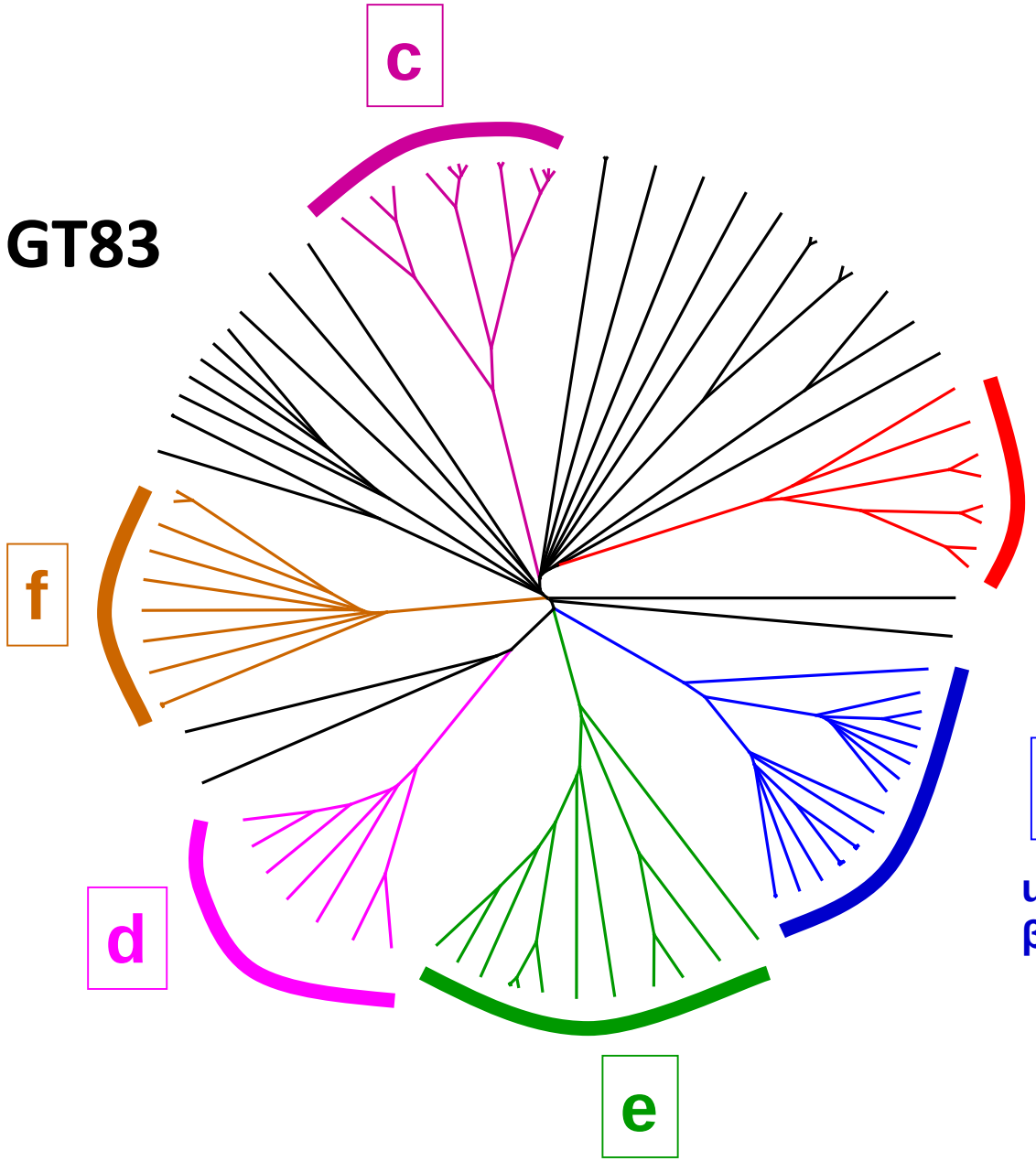
**Should all these be annotated as arabinosyltransferases ?**

**Family GT83**



- **undecaprenyl phosphate-L-Ara4N:4-amino-4-deoxy-β -L-arabinosyltransferase (EC 2.4.2.43)**

# Family GT83




**b** dodecaprenyl-P- β-GalA α-D-galacturonosyl transferase

**a** undecaprenyl phosphate-α-L-Ara4N β-L-Ara4N-transferase

A family of inverting glycosyltransferases using nucleotide monophosphosugar donors

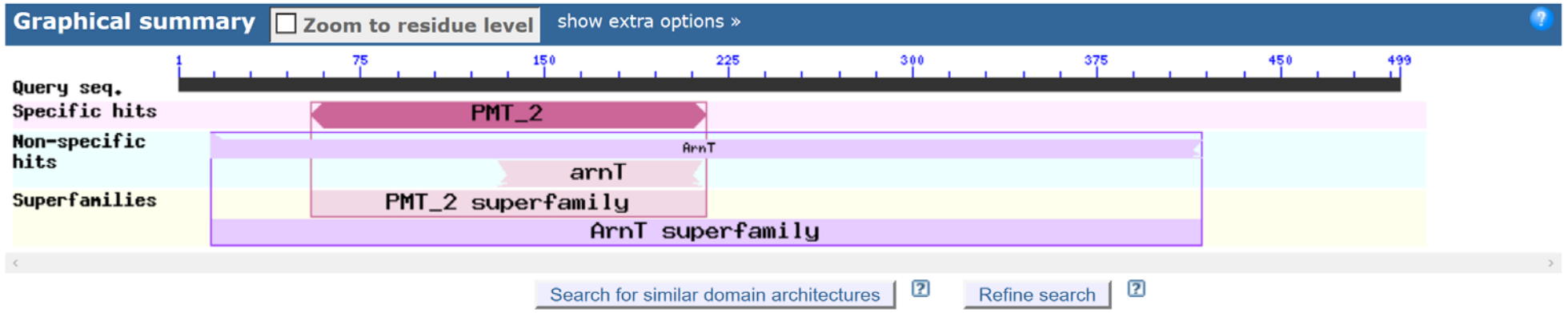
# NCBI nr BLASTp with GT83 dodecaprenyl-P-GalA: LPS core α-D-galacturonosyltransferase from *Rhizobium leguminosarum* (GenBank ABC02169.1)

| Annotation | in top 1000: |
|---|---|
| hypothetical protein | 545 |
| membrane protein | 113 |
| glycosyl transferase | 73 |
| glycosyltransferase family 39 protein | 71 |
| glycosyl transferase family 39 | 65 |
| phospholipid carrier-dependent glycosyltransferase | 31 |
| glycosyltransferase | 29 |
| 4-amino-4-deoxy-L-arabinose transferase | 17 |
| PMT family glycosyltransferase, 4-amino-4-deoxy-L-arabinose transferase | 17 |
| Dolichyl-phosphate-mannose-protein mannosyltransferase | 16 |
| PMT family glycosyltransferase protein | 6 |
| glycosyl transferase family protein | 3 |
| **Lipopolysaccharide core galacturonosyltransferase RgtA** | 3 |
| PMT family glycosyltransferase | 2 |
| putative membrane protein | 2 |
| CAZy families GT83 protein | 1 |
| conserved membrane hypothetical protein | 1 |
| dolichol monophosphate mannose synthase | 1 |
| Glycosyl transferase\|GT83 | 1 |
| glycosyltransferase protein | 1 |
| PMT family glycosyltransferase 4-amino-4-deoxy-L-arabinose transferase | 1 |
| transmembrane protein | 1 |

- **Search done on 23 Feb 2018**
- 22 different annotations in top 1000 results (all significant)
- 3 of 1000 report the correct activity
- But not the sequence I started with
- Correct functional information not properly captured or displayed (not in the top Blast results)
- Wrong functional information tends to percolate more efficiently (Murphy's Law)
- Large diversity of annotation in UniProt
- Swiss-Prot is good now
- No EC number available for this activity

# Protein families servers

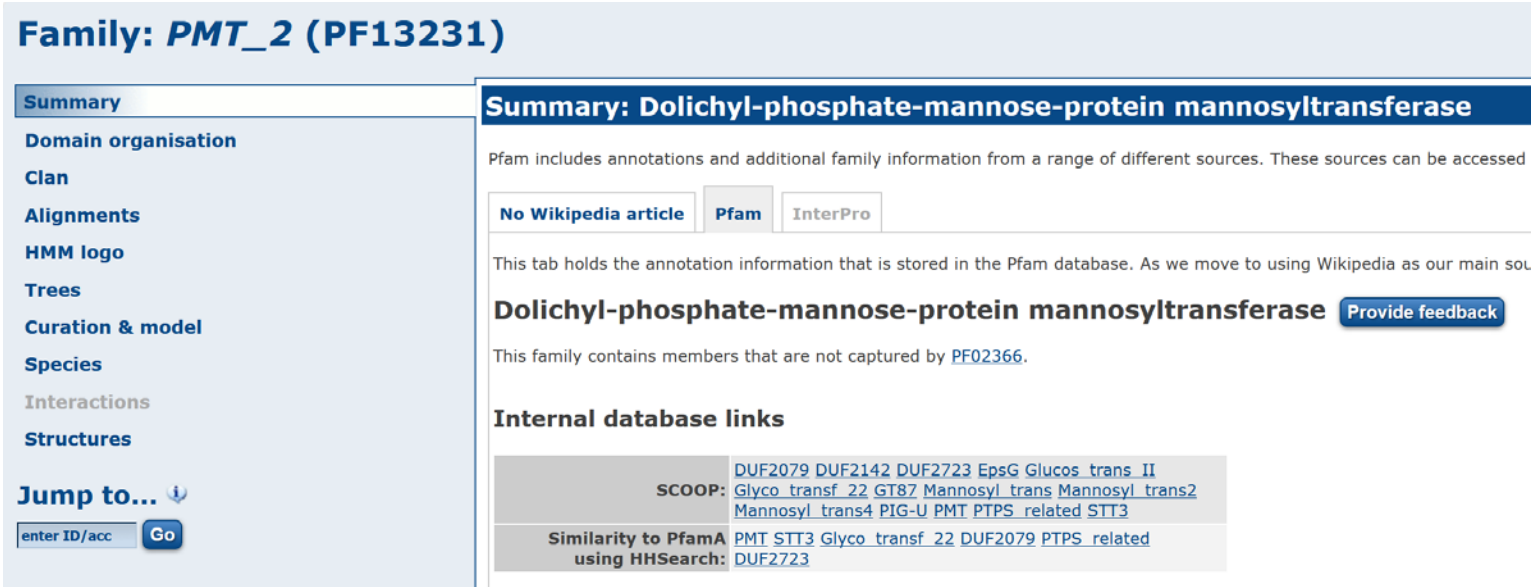
Conserved domains database




**List of domain hits**

| Name | Accession | Description | Interval | E-value |
|---|---|---|---|---|
| [+] ArnT | COG1807 | 4-amino-4-deoxy-L-arabinose transferase or related glycosyltransferase of PMT family [Cell ... | 14-418 | 1.45e-25 |
| [+] PMT_2 | pfam13231 | Dolichyl-phosphate-mannose-protein mannosyltransferase; This family contains members that are ... | 55-216 | 1.68e-15 |
| [+] arnT | PRK13279 | 4-amino-4-deoxy-L-arabinose transferase; Provisional | 131-214 | 2.62e-03 |

Pfam

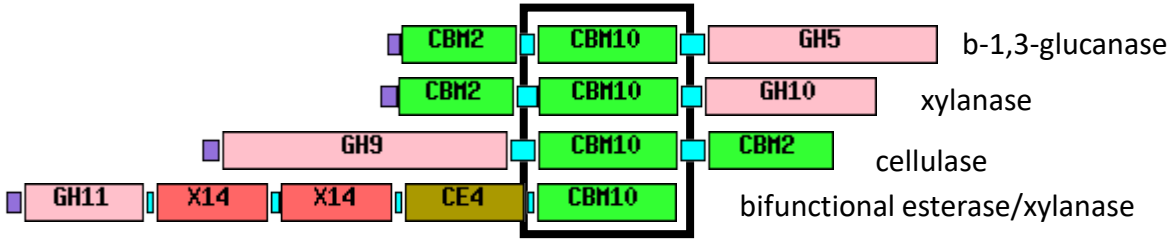


# Where are the problems coming from ?

- Genomes annotated against what is found in general databases **at the moment** they are annotated.
- Subsequent progress in functional biochemistry is rarely (**never**) propagated into "old" genomes. These old, obsolete, genomes then serve for the annotation of the new ones ...
- Errors in general DBs **hard to correct**
- **Modular organization** of CAZymes

# Modularity of CAZymes creates problems for functional annotation

- Best BLAST hit can be on a non catalytic module, making functional prediction hazardous




- What is the function of the X122-containing protein ?

# Other problems

- Human factor (i): tendancy to name families **based on the function** of the first discovered member
- Human factor (ii): it seems that there are as many family definitions and naming conventions as there are scientists
- Lots of putative domains have function-suggesting names (*such as BACON domain: Bacteroidetes-Associated Carbohydrate-binding Often N-terminal*) based on indirect inferrence and no experimental support
- EC numbers / functions assigned to sequences **without biochemistry**

# Problems with the EC numbers

- By definition the EC numbers should be attributed only **after** biochemical characterization
- Unfortunately general databases and genome annotators assign EC numbers **based on sequence similarity**
- For CAZymes, similarity is **frequently too distant** to ensure reliability when passing an EC number by homology
- General databases polluted by **wrong** EC numbers and/or **wrong** function-suggesting names

# How do we cope with these problems in the CAZy database ?

- EC number placed **only** when we have documented evidence for particular activity (literature scans + community of scientists)
- We do **not** transmit **any** functional information by similarity

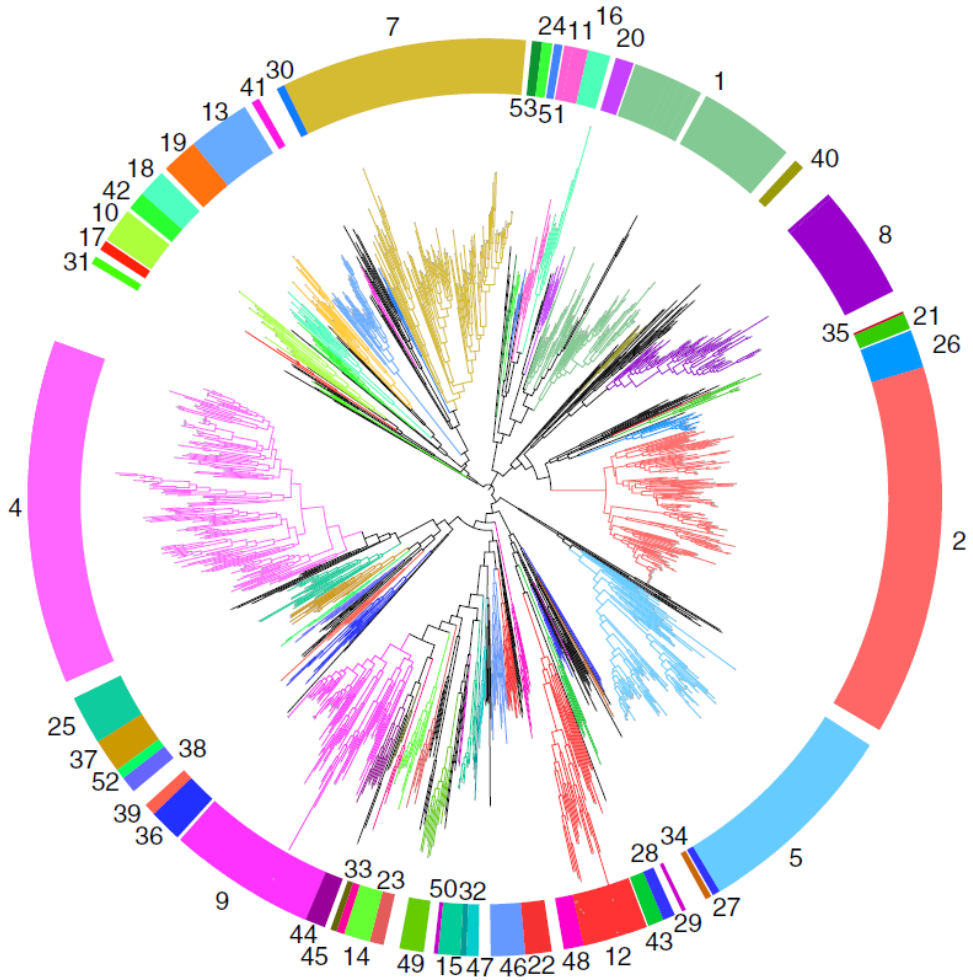
# Practical issues & key challenges

- Capture of functional information:
  - **Revise the EC number system** so that it captures some essential features present in protein families (molecular mechanism, catalytic machinery etc, which can become predictible)
  - Create a world initiative demanding (i) immediately a **simple way** to deposit biochemical characterization dara and (ii) later a more general form (if ever a consensus can be found)
  - A **database of experimentally characterized enzymes** would certainly have great value

# Practical issues & key challenges (continued)

- Many more functions than protein folds: **stop using first function to name a family** as this name will be passed by similarity even remote (especially when the profile is loose to maximize coverage)
- The **full spectrum of activities in a family is rarely known,** some families have only one or a handful characterized members. We certainly know the pitfalls and limitations to function prediction, but those who use our databases do not
- **one cannot use the exact same rules** (thresholds) in different families (variable functional sampling, widely different lengths of the functional modules)
- We must **avoid placing functional predictions in sequence databases**
- Functional predictions should include **explicit integration of the distance** between query and a functionally-characterized sequence

# Subfamilies are currently our best way *en route* to accurate functional prediction

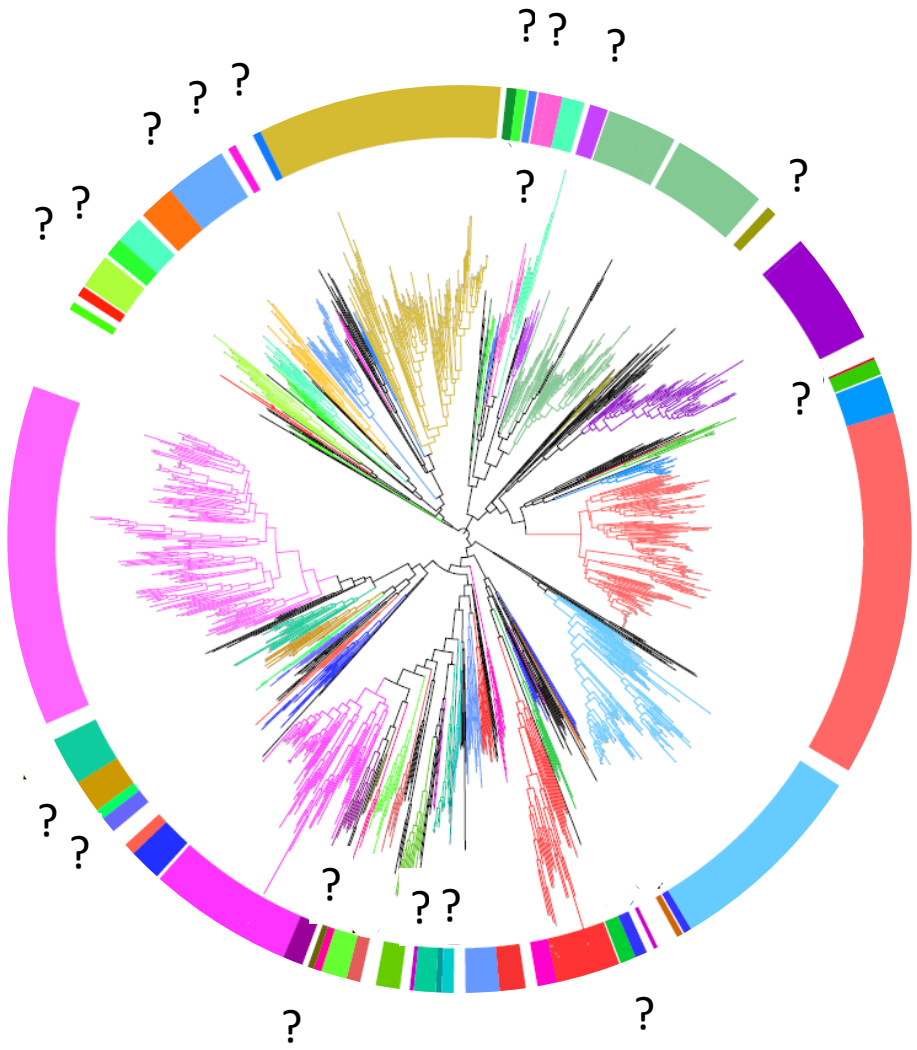



**Family GH5**

- >20 different EC numbers
- 51 subfamilies defined
- 31 subfamilies with known activities
- subfamilies show limited functional variations (→ functional prediction)

Aspeborg et al. 2012

# Subfamilies are currently our best way *en route* to accurate functional prediction



**Family GH5**

- >20 different EC numbers
- 51 subfamilies defined
- 31 subfamilies with known activities
- subfamilies show limited functional variations (→ improved functional prediction)
- 20 subfamilies with no EC number : **→potential for discovery (or redundancy)**

Aspeborg et al. 2012

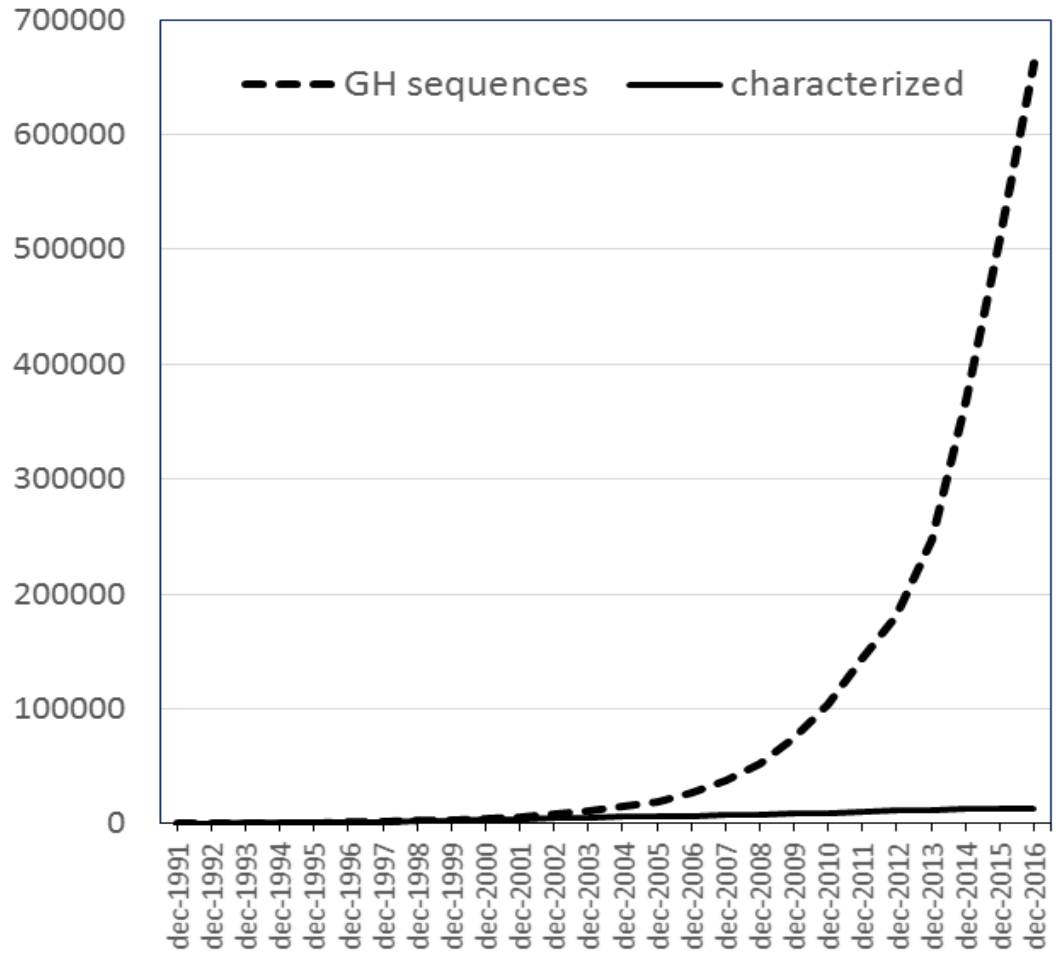
# How can **novel** enzymes be discovered ?

Define a strategy :

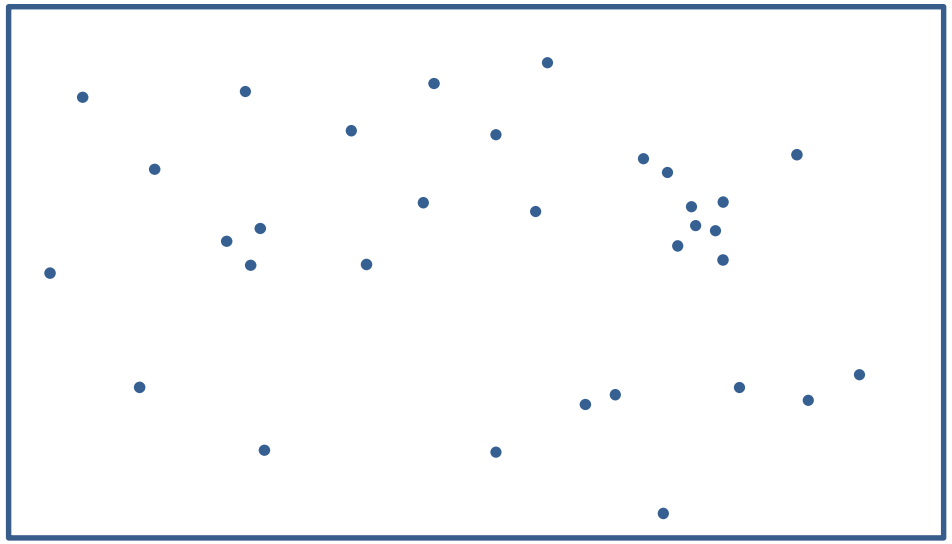
- **chance** (we must be prepared, but we can't count on it; not developped here !)
- **screen** for something you are looking for
- **omics**: secretomics, transcriptomics, genomics
- artificial enzymes; directed evolution towards new reactions
- **bioinformatics** (guilt by association; collaborative networks of enzymes)
- **A systematic exploration** of **nature's** sequence space

# Growth of enzyme sequence data

- Doubling time : 2 years
- Growth of biochemistry **linear**
- A Blast search with a protein sequence will essentially yield sequences that have **not** been characterized
- Immense sequence diversity to explore to **make discoveries** or to make products (in addition to large **redundancy**)



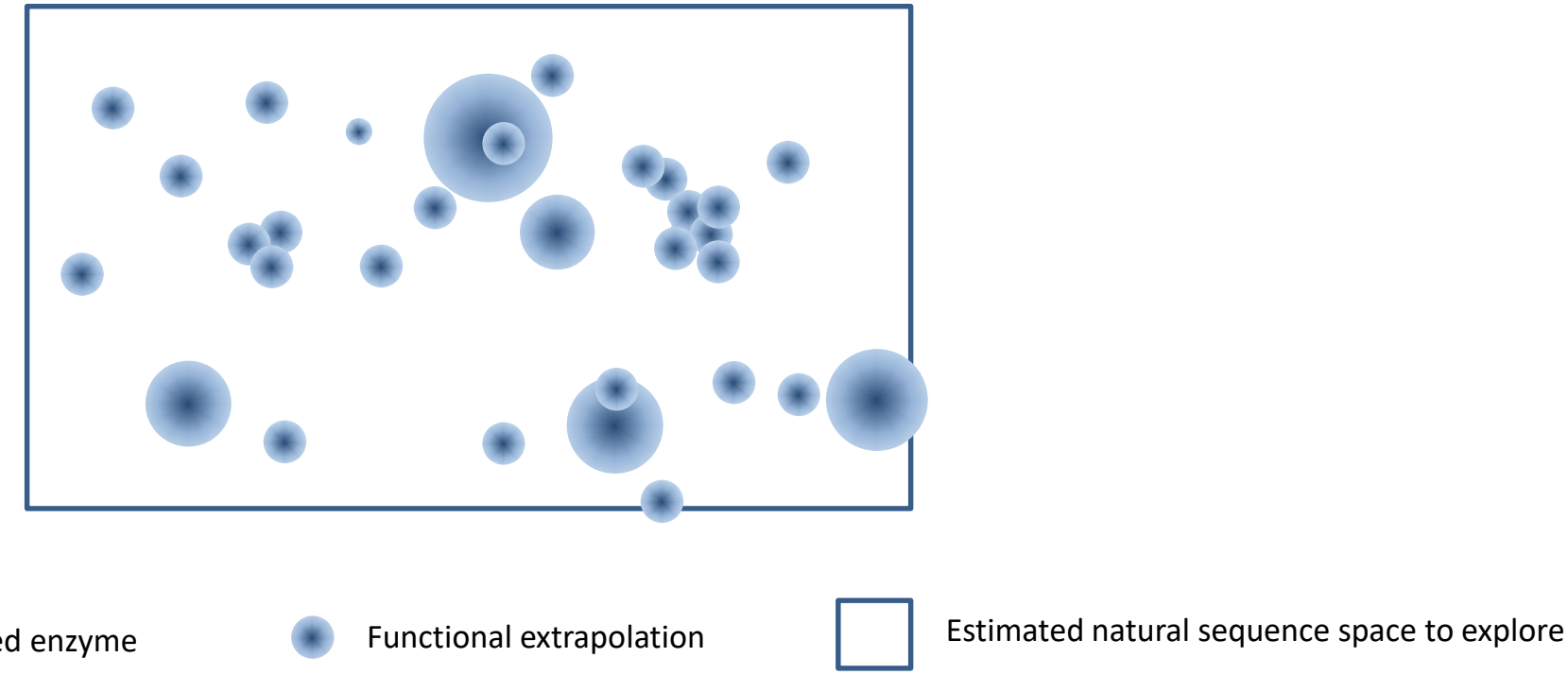
# Systematic exploration of nature's actual sequence space



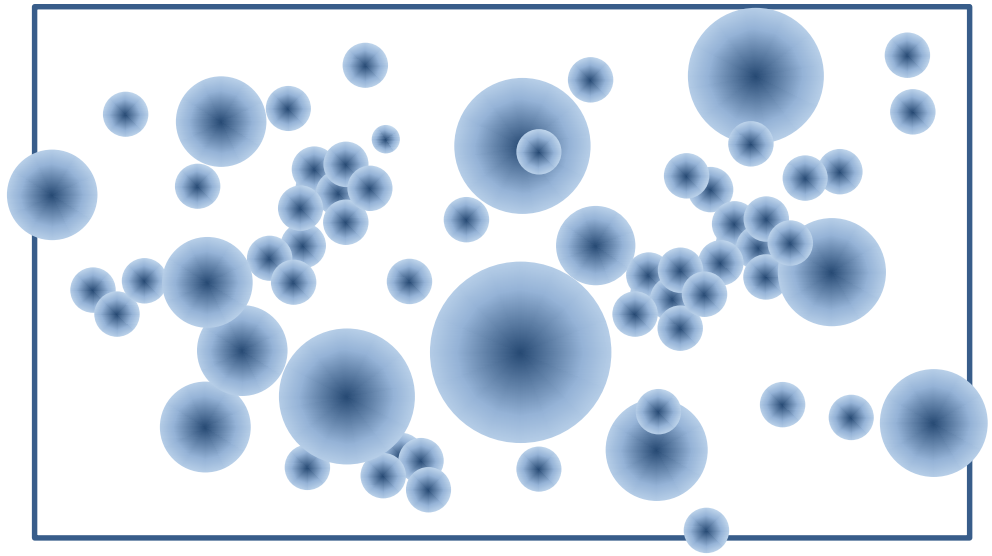
Characterized enzyme

Estimated natural sequence space to explore

# Systematic exploration of nature’s actual sequence space



# Systematic exploration of nature’s actual sequence space

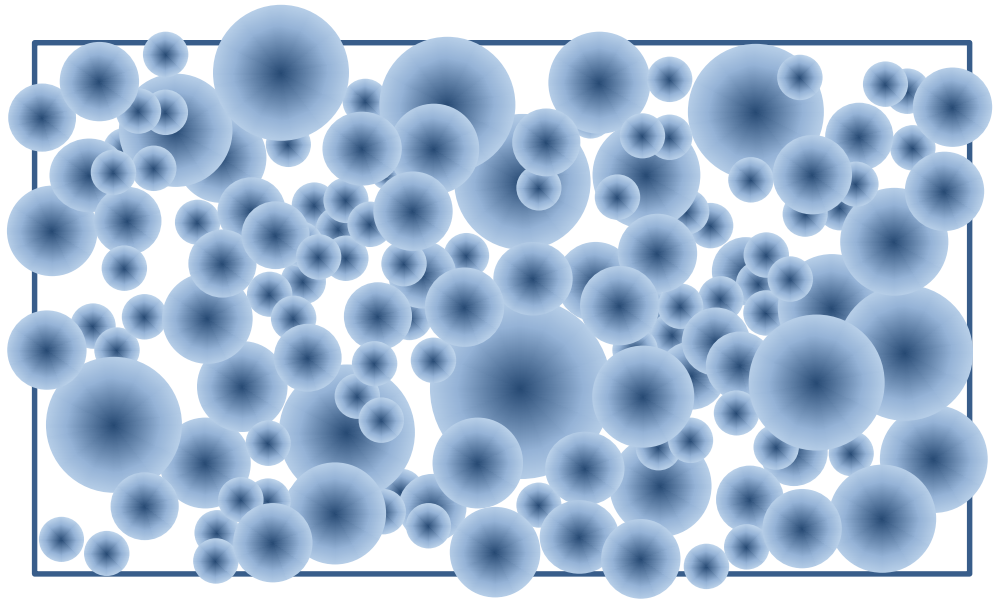


Characterized enzyme

Functional extrapolation

Estimated natural sequence space to explore

# Systematic exploration of nature’s actual sequence space

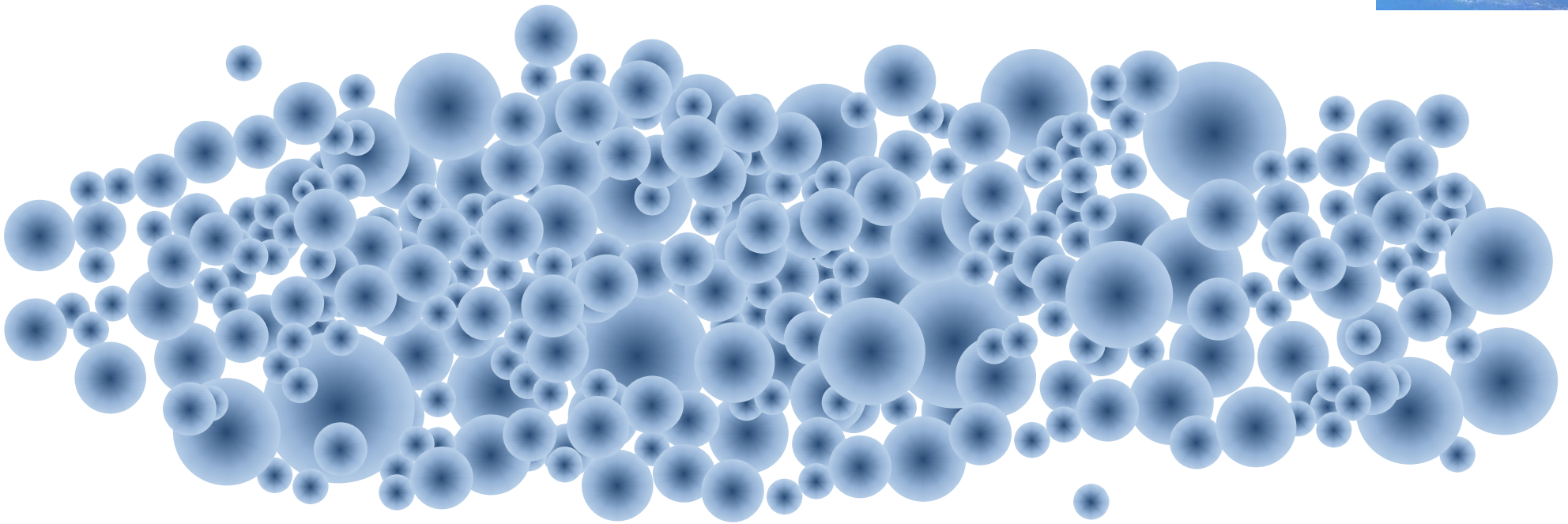


Characterized enzyme

Functional extrapolation

Estimated natural sequence space to explore

# Systematic exploration of nature's actual sequence space



Characterized enzyme

Functional extrapolation

Actual space probably larger than we think

**Nature's sequence diversity provides predictive power**

# Acknowledgments


## AFMB lab Marseille

- **Elodie Drula** (bioinformatics)
- **Marie-Line Garron** (crystallography)
- **Matthieu Hainaut** (genome curation)
- **Pascal Lapébie** (PUL analyses)
- **Nicolas Lenfant** (bioinformatics)
- **Vincent Lombard** (CAZy database)
- **Nicolas Terrapon** (PUL analyses)
- **Renaud Vincentelli** (protein production)

**Postdoc wanted**

## PUL exploration

- **David Bolam** (Newcastle University)
- **Harry J. Gilbert** (Newcastle University)

## Polysaccharide screening

- **William Helbert** (CNRS, Grenoble)