

Bioinformatics Tools for Analyzing Enzyme Families

Greg Butler

Department of Computer Science and Software Engineering

Centre for Structural and Functional Genomics

Concordia University, Montreal

www.cs.concordia.ca/~gregb

gregb@cs.concordia.ca

Outline

Overview of Tools

PipeAlign

Panther

FlowerPower

Case Study on Cellulases

Overview of Tools — Problems Addressed

Multiple Sequence Alignment (MSA)

Problem: Given a set of protein sequences, and an *objective function*, determine the *optimal* alignment of the sequences.

SubProblems: Selecting Homologues to Align; Choice of Objective Function; High-Quality MSA; Removing Outlier Sequences

Phylogenetic Tree Construction

Problem: Given a set of protein sequences, and their *pairwise distances* (using some distance metric), construct a phylogenetic tree.

Classifier for Family — usually Hidden Markov Model (HMM)

Problem: Given a family of protein sequences, and a multiple sequence alignment (MSA) for the family, construct a *classifier* which given a protein sequence can determine whether or not the protein is a member of the family.

Split Family into Subfamilies

Problem: Given a family of protein sequences, and a multiple sequence alignment (MSA) for the family, and ..., determine a *clustering* of the sequences in the family into subfamilies.

Consistency of MSA, Tree, Classifiers for Family and Subfamilies

Overview of Tools

PipeAlign Given a seed sequence, constructs MSA for family and sub-families.

- (optional) include candidate family members

<http://igbmc.u-strasbg.fr/PipeAlign/>

Panther DB of sequences, MSAs, trees, and HMM classifiers for protein families and subfamilies *semi-automatically* for human, mouse, ...

Given query protein sequence, classifiers determine family and subfamily

- can download all HMM classifiers

<https://panther.appliedbiosystems.com/>

FlowerPower Given seed protein sequence, determines the family and subfamily, their MSAs, trees, and HMM classifiers.

- like improved PSI-Blast against UniProt for MSA's
- postprocess MSAs using BÊTE for trees, GTREE for display

http://phylogenomics.berkeley.edu/cgi-bin/flowerpower/input_flowerpower.py

PipeAlign

Ballast

- getting a better set of homologues
- conservation profile

DbClustal

- combined local and global alignment using anchors

NorMD

- a reliable objective function
- normalized Mean Distance scores

RASCAL and LEON

- detection and correction of alignment errors
- removal of outliers
- realignment of blocks and inter-block regions

Secator and DPC

- split into subfamilies

Cellulase Case Study

Dataset of biochemically characterized cellulases for Kwang-Bo Joung:

- 27 endoglucosidases (egl) EC 3.2.1.4
- 23 cellobiohydrolases (cbh) EC 3.2.1.91
- 28 beta-glucosidases (bgl) EC 3.2.1.21

Characterized into 92 families of Glycosol Hydrolases (GH)

Kwang-Bo Joung aligned domains using ClustalW and combined into tree. Used Prosite patterns to clarify subfamily membership.

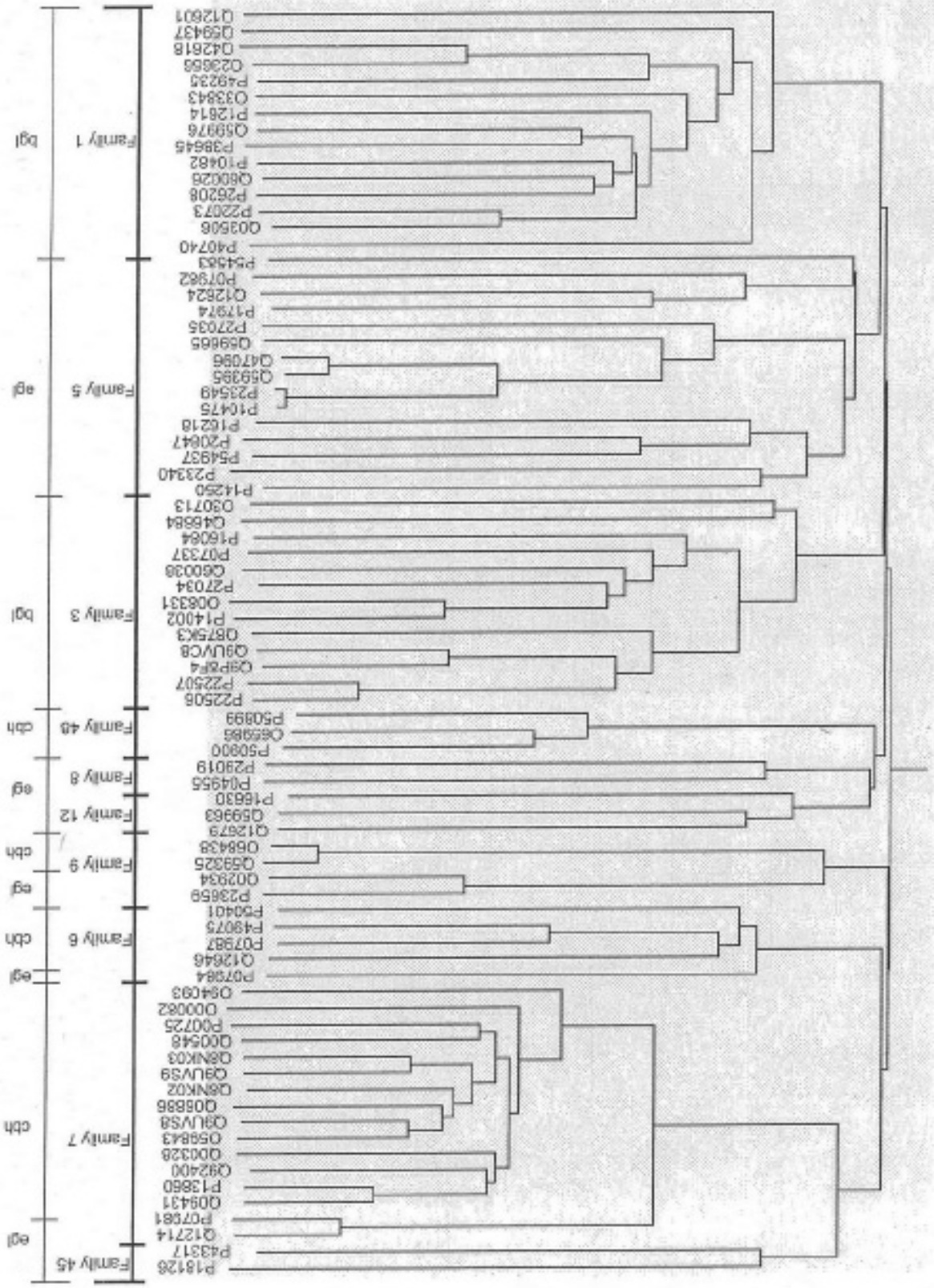
Noted misclassifications:

- P46236, GH Family 6, in SP as *egl*; literature says “cellulase” (ie *egl* or *cbh*); should be *cbh*
- P37698, GH Family 48, should be *cbh* not *egl*

Kwang-Bo Joung’s classification (see tree):

- egl-A of size 4, 2 in GH Family 45, 2 in GH Family 7 (has *cbh*)
- egl-B of size 1 in GH Family 6 (has *cbh*)
- egl-C of size 2 in GH Family 9 (has *cbh*)
- egl-D of size 5 in GH Family 12 and 8
- egl-E of size 15 in GH Family 5
- cbh-A of size 14 in GH Family 7 (has *egl*)
- cbh-B of size 5 in GH Family 6 (has *egl*)
- cbh-C of size 2 in GH Family 9 (has *egl*)
- cbh-D of size 3 in GH Family 48
- bgl-A of size 13 in GH Family 3
- bgl-B of size 15 in GH Family 1

Cellulase Case Study



Cellulase Case Study

Dataset of biochemically characterized cellulases for Kwang-Bo Joung:

- 27 endoglucosidases (egl)
- 23 cellobiohydrolases (cbh)
- 28 beta-glucosidases (bgl)

Automatic analysis by **PipeAlign**

Input: One of the datasets, first sequence as the seed

Parameters allowed PipeAlign to search UniProt and add up to 200 homologues

Results:

egl analysis confirmed several subfamilies in egl-E
— but not P17974 nor Q12624
— and not egl-A to egl-D

cbh analysis confirmed cbh-A
— but not cbh-B to cbh-E

bgl analysis confirmed half of bgl-A

bgl analysis confirmed bgl-B (without P40740)



New session

Paste your **protein** sequence(s) in [FastA format](#) :

Submit pasted sequence

or upload a FastA sequence file :

Submit Uploaded File

PipeAlign

Your reference sequence for the whole PipeAlign process is **eIP14250**.

It will be used as a query for database search.

You should now set the required parameters for every step :

Ballast	<input checked="" type="checkbox"/> Use Filter for BlastP search <input checked="" type="checkbox"/> Perform Blast gapped alignment
DbClustal	Align <input type="text" value="200"/> sequences max. <input type="text" value="200"/> from top Blast/Ballast results <input type="text" value="200"/> with fragments removal + all sequences the sequences you submitted
Rascal	<i>n.a.</i>
NorMD	Score cutoff level: <input type="text" value="6"/> for removal of unrelated or badly-aligned sequences
Clustering method	<input checked="" type="radio"/> Let PipeAlign choose most appropriate clustering method <input type="radio"/> Secator <input type="radio"/> DPC
Launch PipeAlign	Default values

Case Study — Server

Your PipeAlign session has been submitted with the following ID number :

0320-1738-6853

(Please, note this number for future reference)

Case Study — Server

View previous session results

Previous sessions are identified by an ID number in order to let you retrieve their results later on. Please, note that session results are kept only for a week.

Enter previous [session ID](#) :

Case Study — Report for endoglycosidases

PipeAlign 0320-1738-6853 session Report

Clos

Reference sequence: e1P14250

Blast search in PROTEIN database (1,862,423 seq.) reported a **total of 77** sequences:

14 with expect value < **1e-10**

19 with expect value < **1e-06**

27 with expect value < **1e-03**

77 with expect value < **1**

Ballast predicted **20** LMS's covering **231** residues (**35.1%**) of the query (658 aa.).

DbClustal

[47 sequences](#) from the database search results were selected for alignment with the **27** submitted sequences.

The NorMD value for the DbClustal alignment was: **-0.522**

Rascal

After treatment of the DbClustal alignment by Rascal the NorMD value was: **-1.317**

Leon

After treatment of **DbClustal** output by Leon **14** sequences were removed from the alignment

The new NorMD value was: **0.177**

Clustering

Secator distributed the **60** sequences into **5** groups and **1** group of **7** sequences that could not be assigned to any group..

Sequences removed by Leon

e23Q59963 4 0 0.00 7.94 user submitted sequence

e2P17974 1 0 0.00 7.47 user submitted sequence

e18P04955 1 0 0.00 7.81 user submitted sequence

e7P16630 3 0 0.00 7.03 user submitted sequence

e8Q12714 5 0 0.00 8.93 user submitted sequence

e15P43317 5 0 0.00 6.81 user submitted sequence

e19P23659 4 0 0.00 6.98 user submitted sequence

e12P07981 5 0 0.00 9.28 user submitted sequence

e25P29019 2 0 0.00 4.91 user submitted sequence

e21Q02934 2 0 0.00 7.65 user submitted sequence

e24Q12624 1 0 0.00 6.18 user submitted sequence

e4P18126 1 0 0.00 9.25 user submitted sequence

e10P07984 4 0 0.00 7.06 user submitted sequence

e5Q12679 2 0 0.00 9.44 user submitted sequence

Case Study — Report for endoglycosidases

e11P07982				
Q9REW0_ERWCH	[Bacteria] <i>Erwinia chrysanthemi</i> .	Endo-1,4-beta-glucanase (EC 3.2.1.4).		E=0.071
1vjz_A	?	?		E=2e-08
Group 3 : size=7 sequences				
Q9P867_PIREQ	[Eukaryota] <i>Piromyces equi</i> .	Endoglucanase 5A.		E=3e-04
O18454_HETGL	[Eukaryota] <i>Heterodera glycines</i> (Soybean cyst nematode).	Beta-1,4-endoglucanase-2 precursor (EC 3.2.1.4).		E=0.011
O77094_GLORO	[Eukaryota] <i>Globodera rostochiensis</i> (Golden nematode).	Beta-1,4-endoglucanase (EC 3.2.1.4).		E=0.093
Q9GRU5_HETSC	[Eukaryota] <i>Heterodera schachtii</i> .	Beta-1,4-endoglucanase 2 precursor.		E=0.005
Q95UR3_HETGL	[Eukaryota] <i>Heterodera glycines</i> (Soybean cyst nematode).	Beta-1,4-endoglucanase-4.		E=0.042
Q9U6M4_9BILA	[Eukaryota] <i>Globodera tabacum solanacearum</i> .	Beta-1,4-endoglucanase 2 precursor.		E=0.008
O61595_HETGL	[Eukaryota] <i>Heterodera glycines</i> (Soybean cyst nematode).	Beta-1,4-endoglucanase-2.		E=0.005
Group 4 : size=6 sequences				
Q9U6M5_9BILA	[Eukaryota] <i>Globodera tabacum solanacearum</i> .	Beta-1,4-endoglucanase 1 precursor (Fragment).		E=0.032

Case Study: FlowerPower vs PipeAlign

PipeAlign as above on cellobiohydrolases

First sequence as seed, all cbh sequences provided, up to 200 from UniProt

Three families

- cellobiohydrolases (cbh-A) family of size 92
- two families of egl's (???) of size 45 and 43

FlowerPower with first cellobiohydrolase sequence as seed

Two families

- cellobiohydrolases of size 52
- egl-A of size 4

Thank You.

Questions?