Bioinformatic analysis of similarity to allergens

Mgr. Jan Pačes, Ph.D.

Institute of Molecular Genetics, Academy of Sciences, CR

Scope of the work

- Method for allergenicity search used by FAO/WHO
- Analysis of the FARRPd and its impact to the process
- Reflections on E-value cut-off in this context

Allergen

- Defined as a list, not by properties
- Compound (protein, chemical, particle) which causes hypersensitivity in human
- IgE-mediated allergy

- FARRPd: maintained by University of Nebraska
  - http://www.allergenonline.org/
  - Only protein allergens
  - Inclusion of sequence not defined, expert committee opinion
  - Sequences are added and **deleted** yearly

Current recommendations

- WHO: Codex Alimentarius 2009, FAO/WHO 2001
- EFSA: GMO panel 2011
- Method: “The **alignment-based** criterion involving 35 % sequence identity to a known allergen over a window of at least 80 amino acids is considered a minimal requirement”
  - Split protein of interest to 80 aa overlapping strings
  - Make pairwise alignment against allergen database
  - Check for hits longer or equal to 80 aa with identity better or equal to 35%. Gaps are part of the calculation.

Log-odd scores in pairwise alignment

The scores of substitution matrix (with a negative expected value and at least one positive score):

$$\boldsymbol{S_{i,j}} = \left(\frac{\boldsymbol{q_{i,j}}}{\boldsymbol{p_i p_j}}\right)/\boldsymbol{\lambda}$$

- where λ is a positive *scale parameter*
- $q_{i,j}$ are positive numbers that sum to 1
- $p_i p_j$ are background frequencies

Karlin, Altschul, PNAS 1990

Expected number of high-scoring alignments

For two random seqences *m* and *n* the *expected number* of alignments better than score *s* is:

$$\mathbf{E = K}\ \boldsymbol{mn\ e^{-\lambda s}}$$

Where **<u>K</u>** is a calculable positive parameter dependent on **<u>substitution matrix and background letter frequencies</u>**. This is called *E-value* associated with score *S*.

The number of such high-scoring alignments shows *Poisson* distribution. Thus the probability of finding *at least one* alignment with score > *S* is:

$$\boldsymbol{p = 1 - e^{-E}}$$

Karlin, Altschul, PNAS 1990

E-value threshold

Sources of possible problems:

- Specificity vs. Sensitivity
  - Specificity = TruePositive / (TruePositive + FalsePositive)
  - Sensitivity = TruePositive / (TruePositive + FalseNegative)
- Method for similarity search is based on evolutionary and biochemical similarity (BLOSUM, PAM similarity matrices), not on structural similarity
- FARRPd has special properties, which affect E-value calculation
  - uneven distribution of entries
  - limited cover of protein realm

“Hits” in common genomes

Organism	Alignments >35%/80aa	Known allergens	Genome size	frequency
maize	20,326	24	~5 Gbp	~ 1/250 kbp
soybean	17,011	38	~1.1 Gbp	~ 1/65 kbp
rice	27,669	20	~.4 Gbp	~ 1/15 kbp

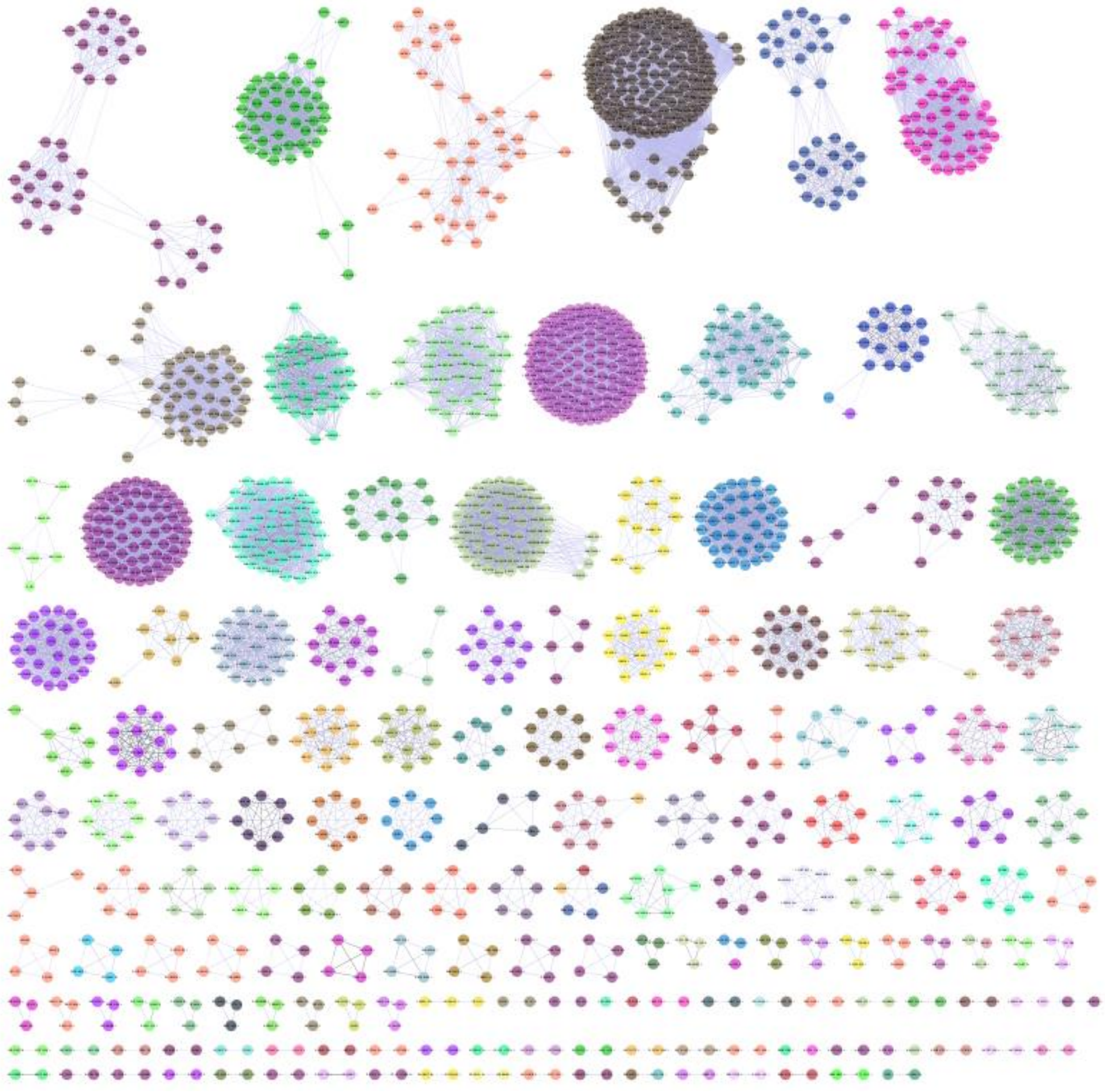
based on Harper et al. 2012; Young et al. 2012

Clusters in FARRPd

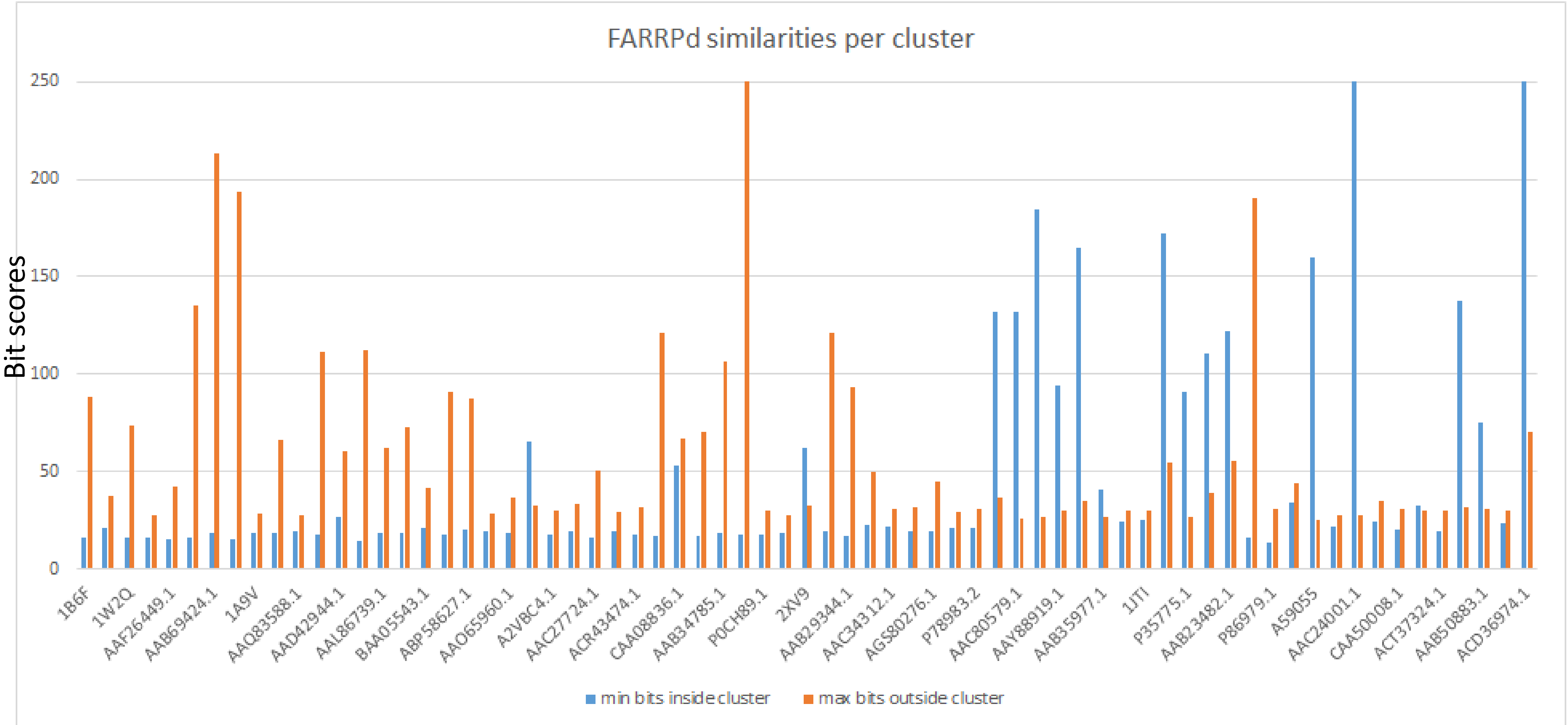
Each color represents one cluster of allergens (allergens of the same group)

Procedure:

- Complete network of all similarities in FARRPd (FASTA)
- Clustering by MCL algorithm based on bit-scores
- Edge-weighted organic layout by Cytoscape



Minimal score inside cluster vs. best score outside cluster



Real example: Maximal E-value so far

```
command line:fasta -w 80 -m 9 -C 20 -Q -3 -E1000 -d1000 low_hit_example.fna farrp_v15.fasta >low_hit_example-hits.txt
>>gi|27806257|ref|NP_776945.1| collagen alpha-2(I) chain precursor [Bos taurus] (1364 aa)
initn:  60 init1:  39 opt:  47  Z-score: 53.3  bits: 18.2 E(1897): 5.8e+02 Smith-Waterman score: 94; 37.6%
identity (55.3% similar) in 85 aa overlap (1-79:261-341)

                              10           20        30        40         50        60        70
test_allergen       PEPRLSP--SPG-VGRGGVRRV--LEPRPSPSPGVGRWGIRRLPEDRLSPS-PGVGRGVIRRLPEPRPSPSPRVGWGGASYGARG
                    :. .:.:  .:: .: .: :    :    .:    :  :   ::  . . . :::. :.   :: ::  :.: :: .::. ::::
gi|27806257|ref|NP_ PKGELGPVGNPGPAGPAGPRGEVGLPGLSGPVGPPGNPGANGLPGAKGAAGLPGVA-GAPG-LPGPRGIPGP-VGAAGAT-GARG
```

Resources

- FARRPd http://www.allergenonline.com/
    - http://farrp.unl.edu/resources/farrp-databases
    - peer reviewed allergen list and sequence searchable database
- FASTA, BLAST pairwise alignment programs
- MCL http://micans.org/mcl/
    - Markov Cluster Algorithm, a fast and scalable unsupervised cluster algorithm for graphs (networks)
    - Enright A.J., Van Dongen S., Ouzounis C.A. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Research 30(7):1575-1584 (2002).
- Cytoscape http://www.cytoscape.org/
    - Graph visualization

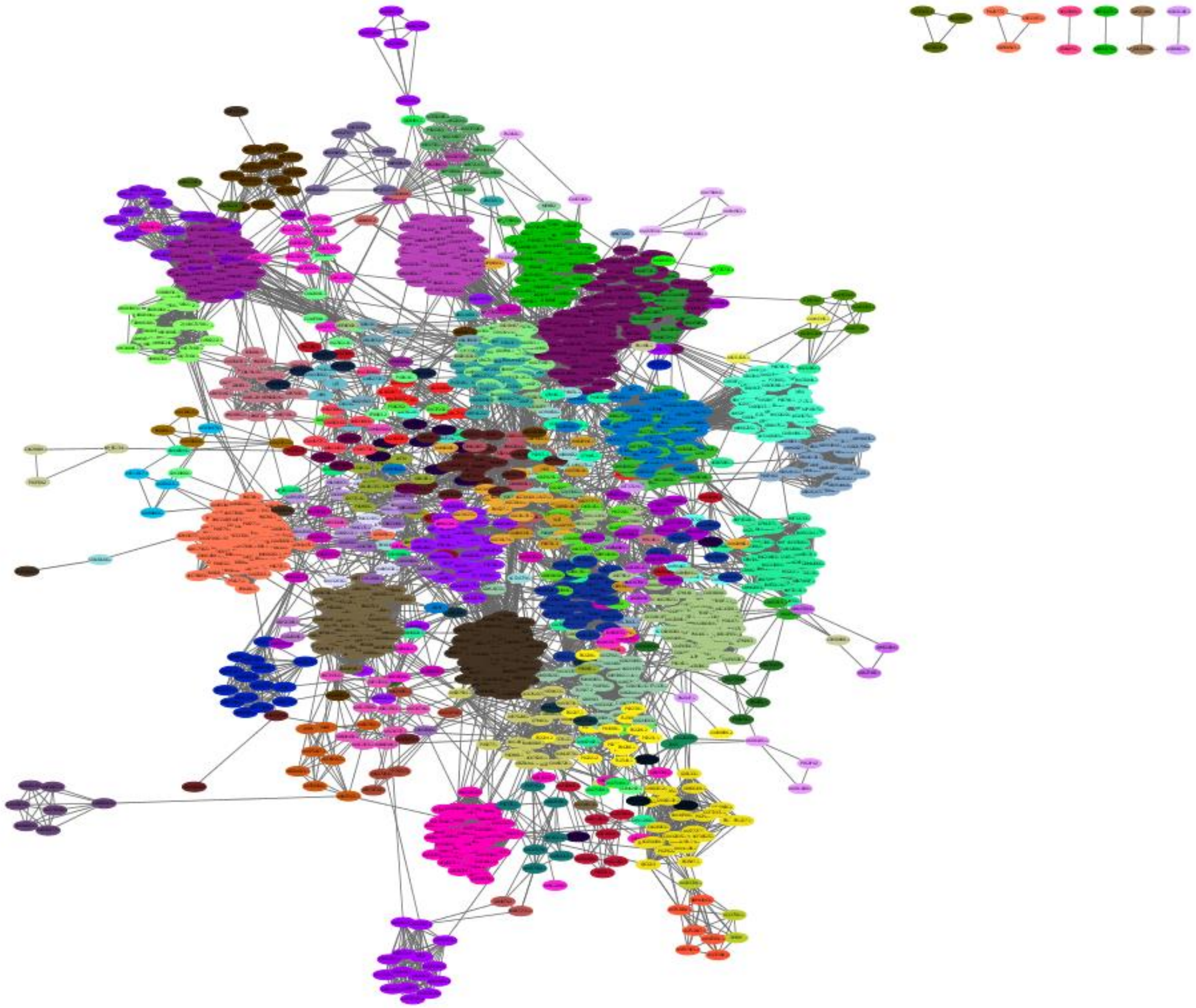
Acknowledgement

Institute of Molecular Genetics, AS CR

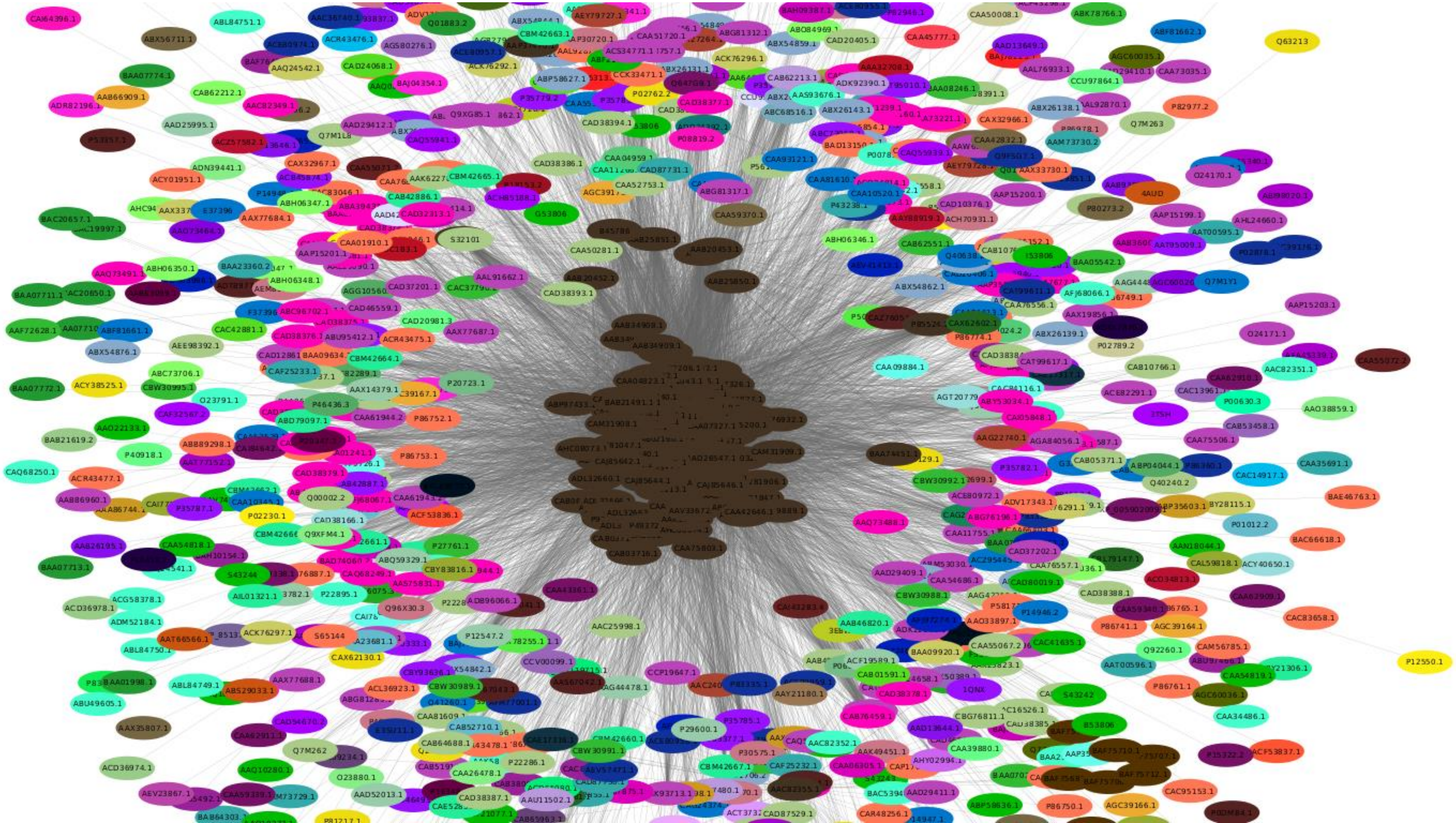
ELIXIR European Bioinformatics Infrastructure

EFSA

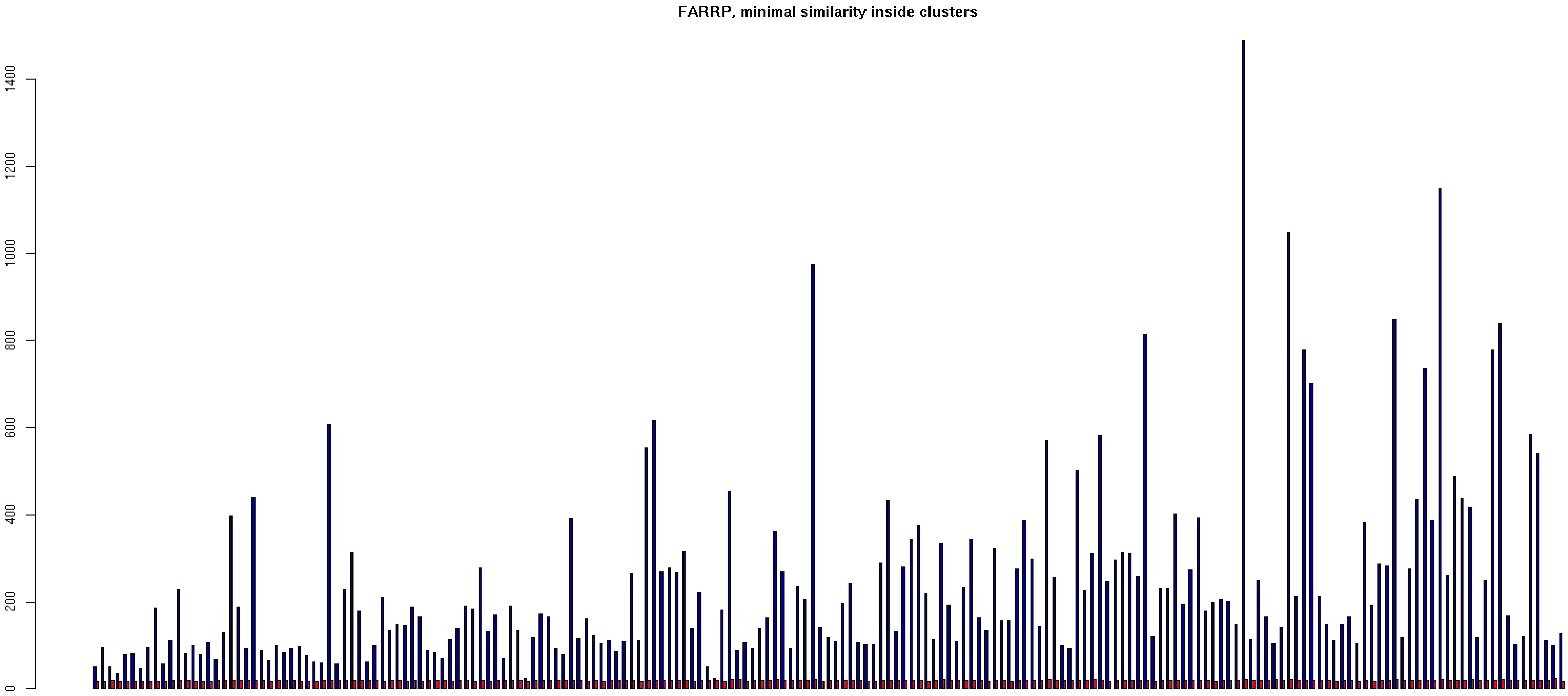
Full FARRPd similarity network



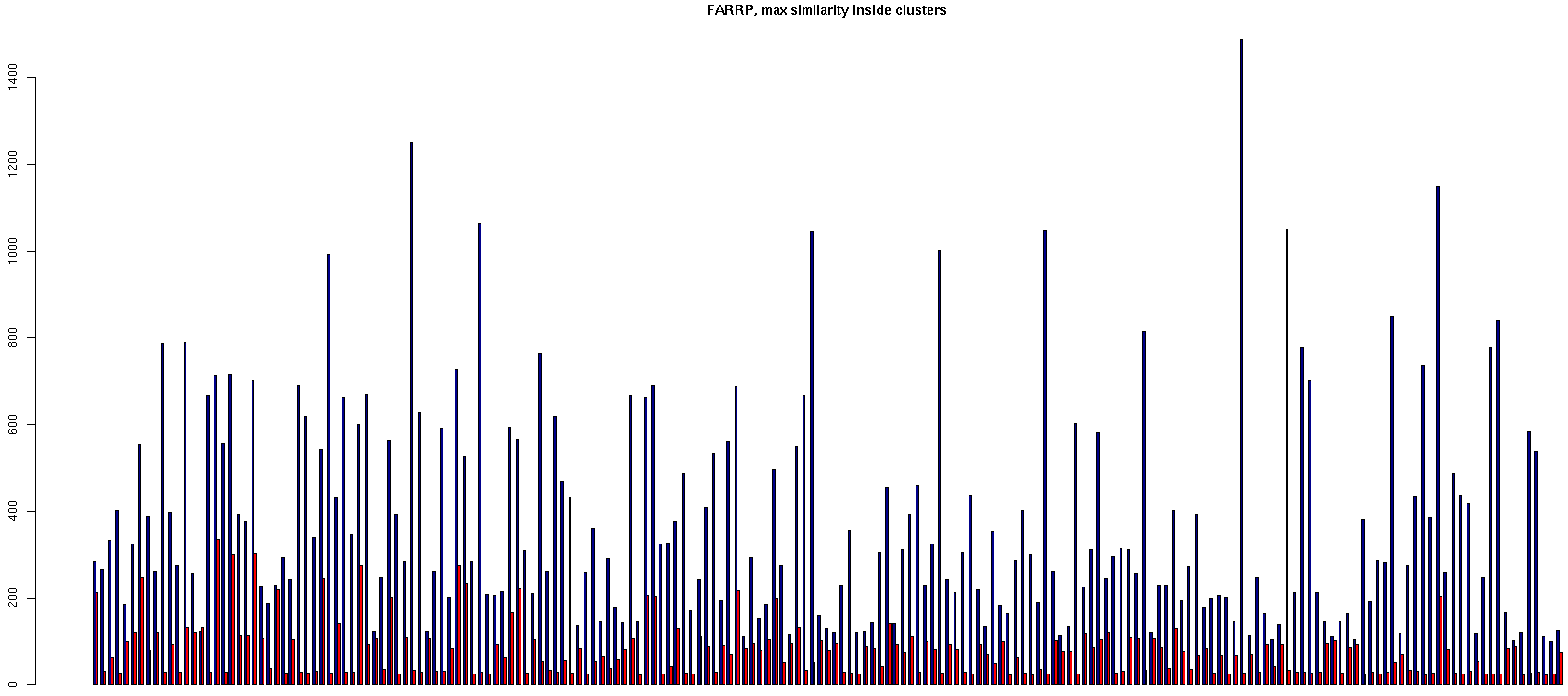
Mapping outside cluster



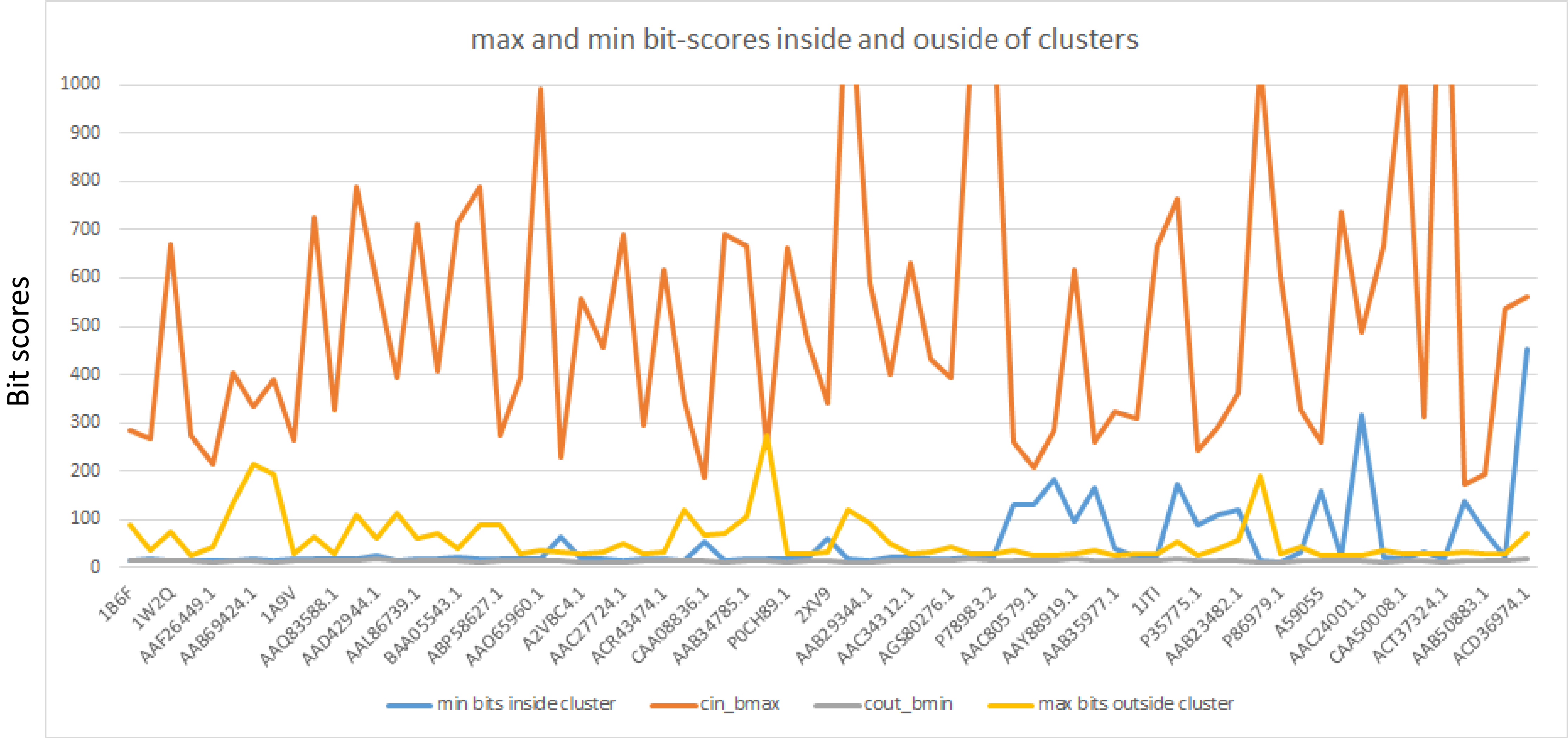
MIN


FARRP, minimal similarity inside clusters
0
200
400
600
800
1000
1200
1400

MAX


FARRP, max similarity inside clusters
0
200
400
600
800
1000
1200
1400

Score inside and outside clusters


max and min bit-scores inside and ouside of clusters
Bit scores
1B6F
1W2Q
AAF26449.1
AAB69424.1
1A9V
AAQ83588.1
AAD42944.1
AAL86739.1
BAA05543.1
ABP58627.1
AAO65960.1
A2VBC4.1
AAC27724.1
ACR43474.1
CAA08836.1
AAB34785.1
P0CH89.1
2XV9
AAB29344.1
AAC34312.1
AGS80276.1
P78983.2
AAC80579.1
AAY88919.1
AAB35977.1
1JTI
P35775.1
AAB23482.1
P86979.1
A59055
AAC24001.1
CAA50008.1
ACT37324.1
AAB50883.1
ACD36974.1
min bits inside cluster
cin_bmax
cout_bmin
max bits outside cluster