Application of the Lempel-Ziv complexity to the alignment-free sequence comparison of protein families

Sofiène Bacha and Denis Baurain

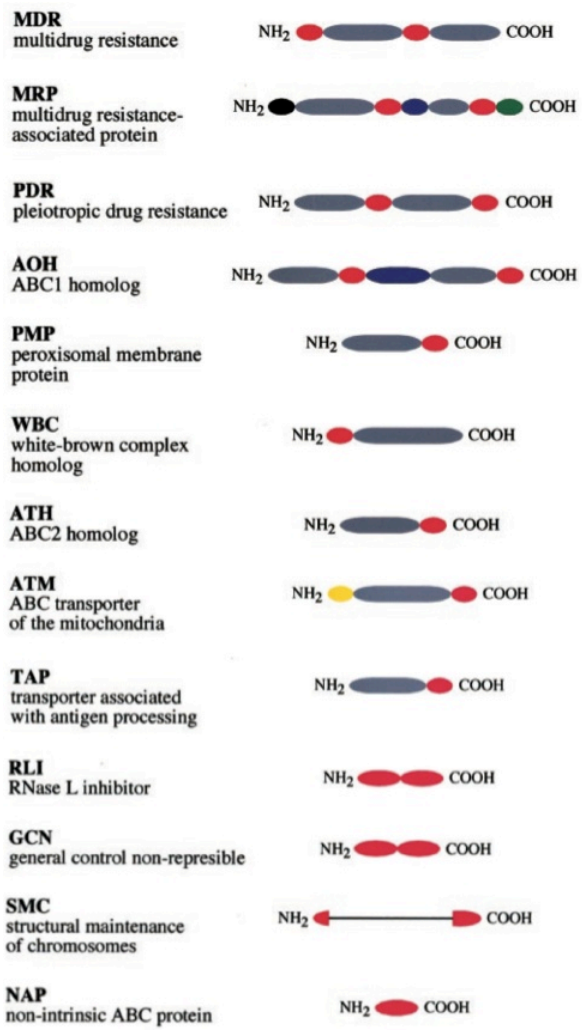
Université de Liège, Belgium

- **Background**
  - Issues with multiple alignment
  - Alignment-free sequence comparison
  - Lempel-Ziv complexity
  - LZ distance metrics
- **Methods**
  - Encoding schemes
  - Benchmarking strategy
- **Results**
  - Benchmark graphs
  - Addendum: decision trees
  - Application to phylogenetics
- **Discussion**
  - Performance considerations
  - Conclusions and perspectives

Background

Issues with multiple alignment

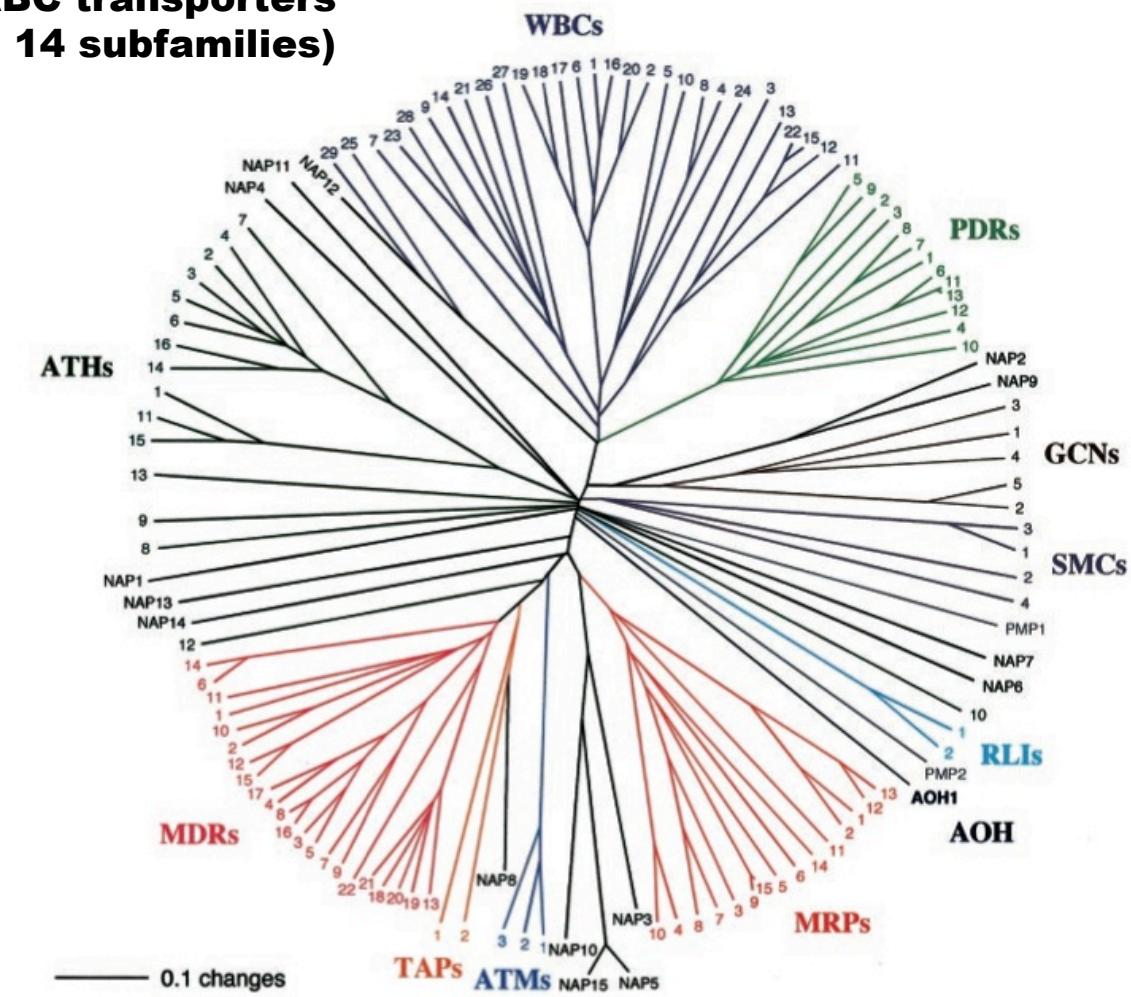
- **multiprotein family phylogenies are *NOT* organism phylogenies**
- **paralogues vs orthologues**
- **paralogues often exhibit domain shuffling and/or domain duplication**
- **rapid diversification may follow gene duplication and then give rise to numerous subfamilies (e.g. ABC transporters)**



Sanchez-Fernandez *et al.* (2001) *J. Biol. Chem.*, **276**, 30231-30244

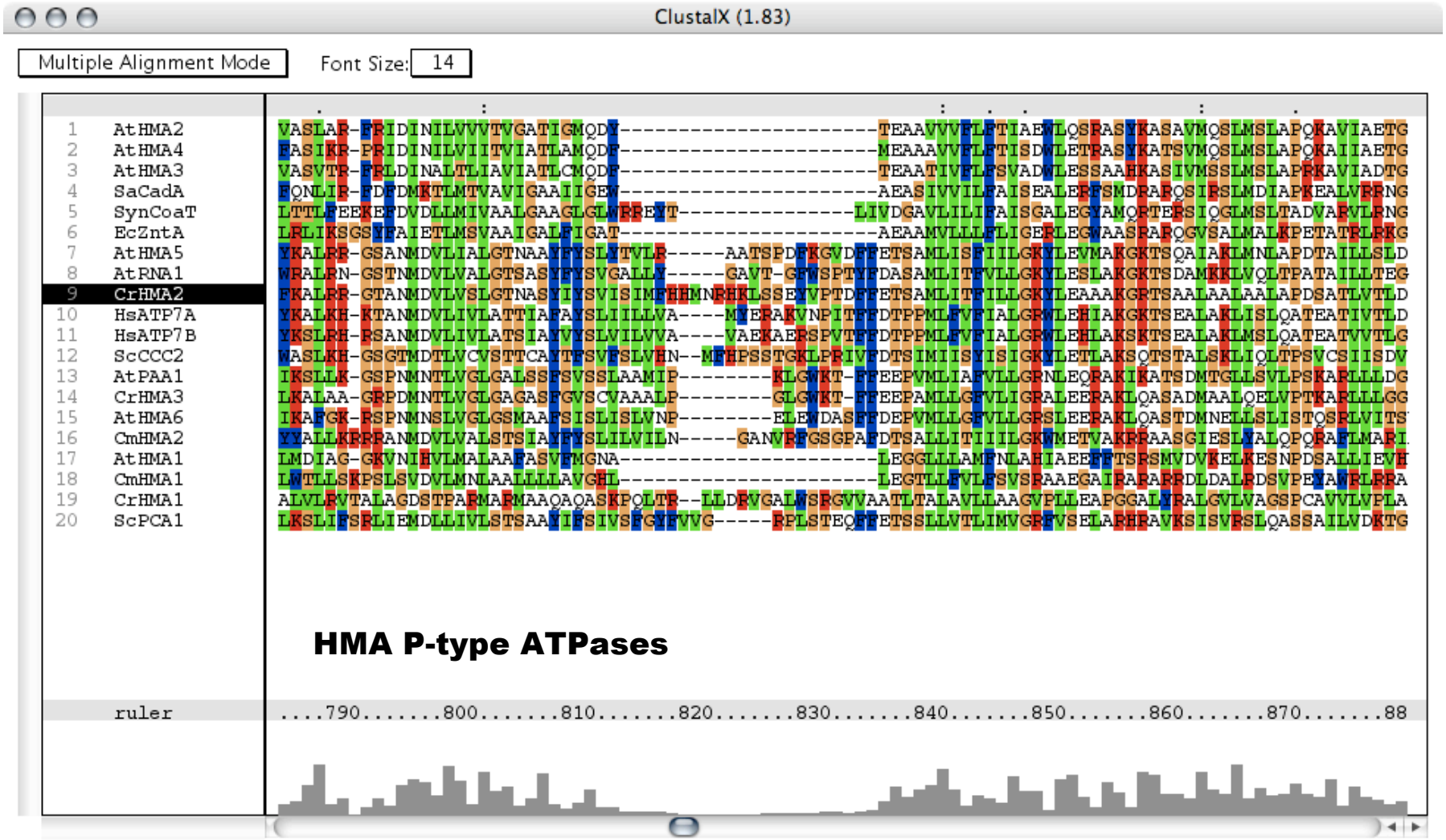
Issues with multiple alignment

***Arabidopsis* ABC transporters**
(129 proteins, 14 subfamilies)



Sanchez-Fernandez *et al.* (2001) *J. Biol. Chem.*, **276**, 30231-30244

Issues with multiple alignment



Done.

Hanikenne *et al.* (2005) *Plant Physiol.*, **137**, 428-446

Alignement-free sequence comparison

- **word statistics**
  - **vectors of counts or frequencies of *k*-mers**
  - **e.g. squared Euclidean distance:**

$$d_k^E(S,Q) = \sum_{i=1}^{K} (c_{k,i}^S - c_{k,i}^Q)^2$$

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
seq S	3	2	1	2	3	7	3	3	5	8	5	3	5	2	9	4	3	4	2	3
seq Q	4	2	1	2	2	6	3	2	6	8	5	3	7	2	6	4	4	3	2	4

k=1; K=20

- **information theory**
  - **algorithmic complexity**
  - **estimation through sequence compression**

Vinga and Almeida (2003) *Bioinformatics*, **19**, 513-523

Lempel-Ziv complexity

- **Exhaustive history H_E**
  - $S = AACGTACCATTG$
  - $H_E(S) = A \bullet AC \bullet G \bullet T \bullet ACC \bullet AT \bullet TG$
  - $Q = ACGGTCACCAA$
  - $H_E(Q) = A \bullet C \bullet G \bullet GT \bullet CA \bullet CC \bullet AA$
- **The LZ complexity is the number of components in H_E**
  - $c(S) = c_E(S) = 7$
  - $c(Q) = 7$
- **Subadditivity of the LZ complexity**
  - $SQ = AACGTACCATTGACGGTCACCAA$
  - $H_E(SQ) = A \bullet AC \bullet G \bullet T \bullet ACC \bullet AT \bullet TG \bullet ACGG \bullet TC \bullet ACCAA;\ c(SQ) = 10$
  - $c(SQ) \leq c(S) + c(Q)$ [indeed, $10 < 7 + 7$]
- **LZ complexity and sequence similarity**
  - $R = CTAGGGACTTAT$
  - $H_E(R) = C \bullet T \bullet A \bullet G \bullet GGA \bullet CTT \bullet AT;\ c(R)=7$
  - $H_E(RQ) = C \bullet T \bullet A \bullet G \bullet GGA \bullet CTT \bullet AT \bullet ACG \bullet GT \bullet CA \bullet CC \bullet AA;\ c(RQ) = 12$
  - $c(SQ) < c(RQ)$ because S is closer [more similar] to Q than to R [e.g. ACG and $ACCA$]

Lempel and Ziv (1976) *IEEE Trans. Inform. Theory*, **IT-22**, 75-81

LZ distance metrics

$$d(S,Q) = max\{c(SQ) - c(S), c(QS) - c(Q)\} \quad (1)$$

$$d^*(S,Q) = \frac{max\{c(SQ) - c(S), c(QS) - c(Q)\}}{max\{c(S), c(Q)\}} \quad (2)$$

$$d_1(S,Q) = c(SQ) - c(S) + c(QS) - c(Q) \quad (3)$$

$$d_1^*(S,Q) = \frac{c(SQ) - c(S) + c(QS) - c(Q)}{c(SQ)} \quad (4)$$

$$d_1^{**}(S,Q) = \frac{c(SQ) - c(S) + c(QS) - c(Q)}{\frac{1}{2}[c(SQ) + c(QS)]} \quad (5)$$

Otu and Sayood (2003) *Bioinformatics*, **19**, 2122-2130

Methods

Encoding schemes

- **DNA sequences**
  - **the *exact match* approach of the LZ distance metrics works well with the small and simple DNA alphabet**
- **AA sequences**
  - **LZ dm are expected to miss the subtle and overlapping *similarities* characterizing the larger and more complex AA-alphabet**
  - ***substitution matrices* (e.g. PAM, BLOSUM…) are not applicable since LZ dm does not compare residues on a *pairwise* basis**
- **strategy**
  - **we propose several variants of a simple approach where AA sequences are *encoded* to different alphabets *prior to the computation* of the LZ complexity**
- **key idea**
  - ***to capture as much information as possible* in order to enhance the alignment-free sequence comparison of proteins**

Encoding schemes

1. binary codons

	T	C	A	G	T	C	A	G	T	C	A	G
Met (M)	0	0	1	0	1	0	0	0	0	0	0	1
Cys (C)	1	0	0	0	0	0	0	1	1	1	0	0
Ile (I)	0	0	1	0	1	0	0	0	1	1	1	0
Pro (P)	0	1	0	0	0	1	0	0	1	1	1	1
Arg (R)	0	1	1	0	0	0	0	1	1	1	1	1

12 bytes

2. binary codons; 3rd position discarded

	T	C	A	G	T	C	A	G
Met (M)	0	0	1	0	1	0	0	0
Cys (C)	1	0	0	0	0	0	0	1
Ile (I)	0	0	1	0	1	0	0	0
Pro (P)	0	1	0	0	0	1	0	0
Arg (R)	0	1	1	0	0	0	0	1

8 bytes

4. alphanumeric codons

Met (M)	A	T	G															
Cys (C)	T	G	T	T	G	C												
Ile (I)	A	T	T	A	T	C	A	T	A									
Pro (P)	C	C	T	C	C	C	C	C	A	C	C	G						
Arg (R)	C	G	T	C	G	C	C	G	A	C	G	G	A	G	A	A	G	G

3 to 18 bytes

5. alphanumeric codons; compressed

Met (M)	A	T	G				
Cys (C)	T	G	T	C			
Ile (I)	A	T	T	C	A		
Pro (P)	C	C	T	C	A	G	
Arg (R)	C	A	G	T	C	A	G
Ser (S)	T	A	C	G	T	C	A	G

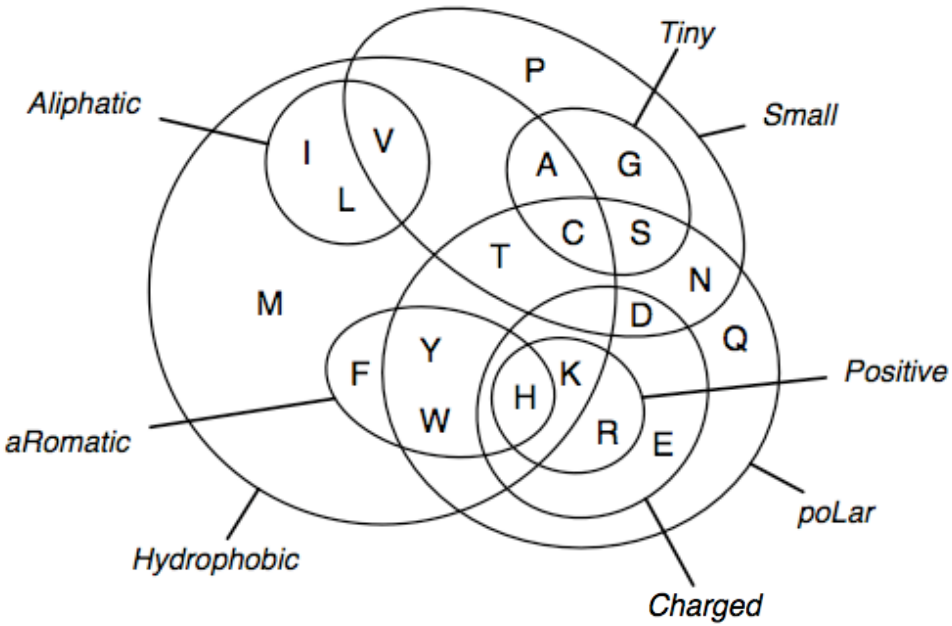
3 to 8 bytes

Encoding schemes

3. biochemical groupings; binary sets

	T	S	P	L	C	H	R	A
Met (M)	0	0	0	0	0	1	0	0
Cys (C)	1	1	0	1	0	1	0	0
Ile (I)	0	0	0	0	0	1	0	1
Pro (P)	0	1	0	0	0	0	0	0
Arg (R)	0	0	1	1	1	0	0	0

8 bytes



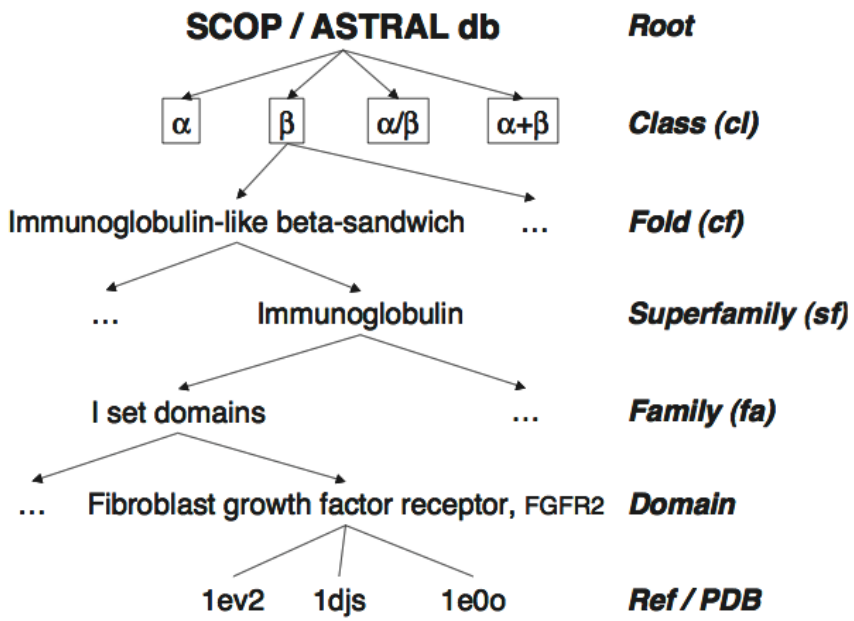
6. biochemical groupings; single byte

Ala (A)							
Gly (G)							
Ile (I)	Phe (F)	His (H)					
Leu (L)	Trp (W)	Lys (K)	Asp (D)	Asn (N)	Ser (S)	Cys (C)	
Val (V)	Tyr (Y)	Arg (R)	Glu (E)	Gln (Q)	Thr (T)	Met (M)	Pro (P)
aLiphatic	*aRomatic*	*Basic*	*Acidic*	*aMide*	*alCohol*	*Sulfur*	*Proline*
L	R	B	A	M	C	S	P

1 byte

Benchmarking strategy

- **dataset**
  - **1,683 protein domains**
- **SCOP / ASTRAL db**
  - **hierarchical organization based on 3D-structures**
  - ***family* and *superfamily* levels reflect phylogeny**
  - ***fold* and *class* levels reflect broad structure similarity**
- **distance methods**
  - **squared Euclidean distance**
  - **W-metric (BLOSUM50)**
  - **SW local alignment score**
  - **LZ dm (raw and encoded)**
- **ROC curves and AUCs**

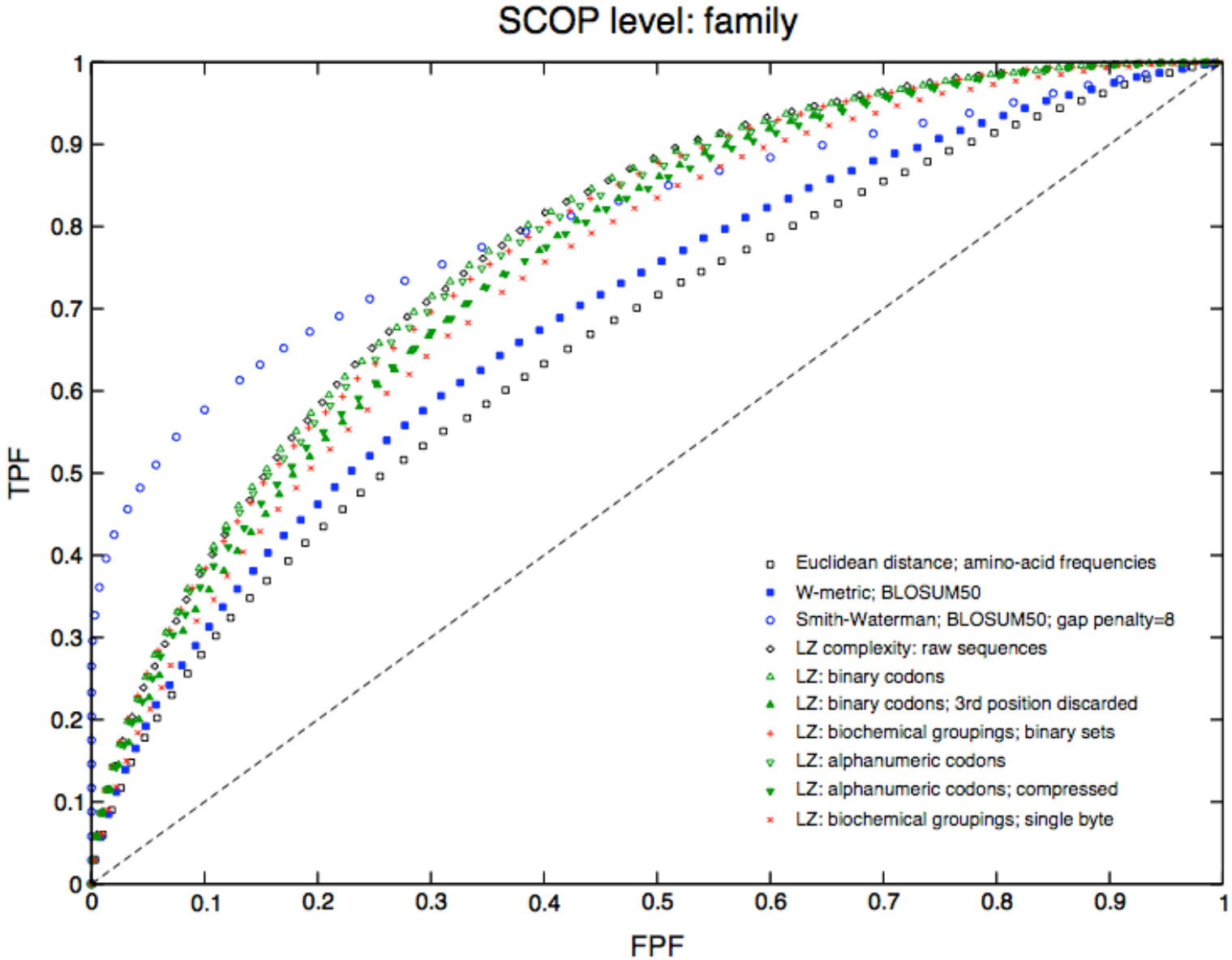


$$d^W(S,Q) = \sum_{i=1}^{K}\sum_{j=1}^{K}(f_i^S - f_i^Q)\cdot(f_j^S - f_j^Q)\cdot w_{ij}$$

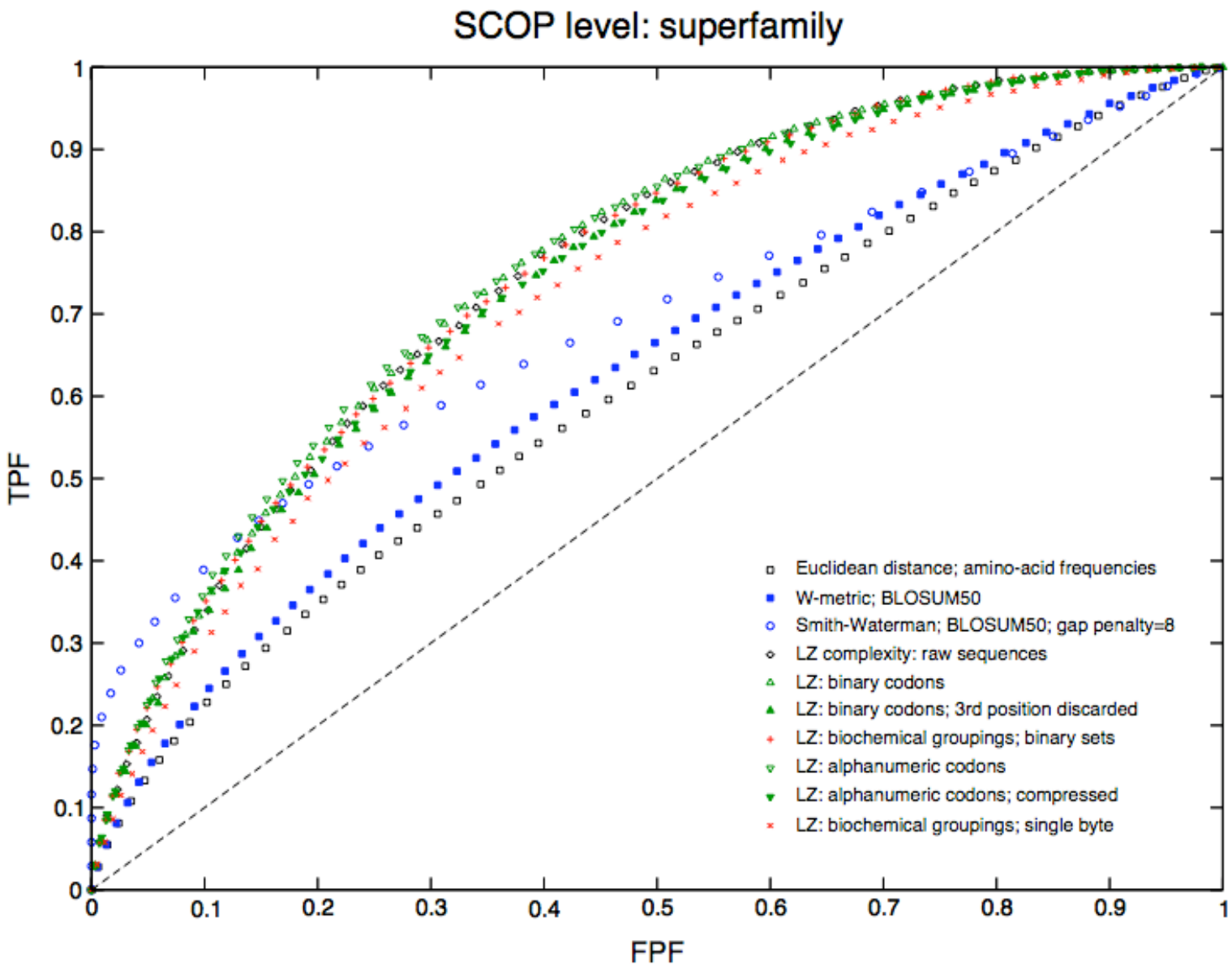
Vinga *et al.* (2004) *Bioinformatics*, **20**, 206-215

Results

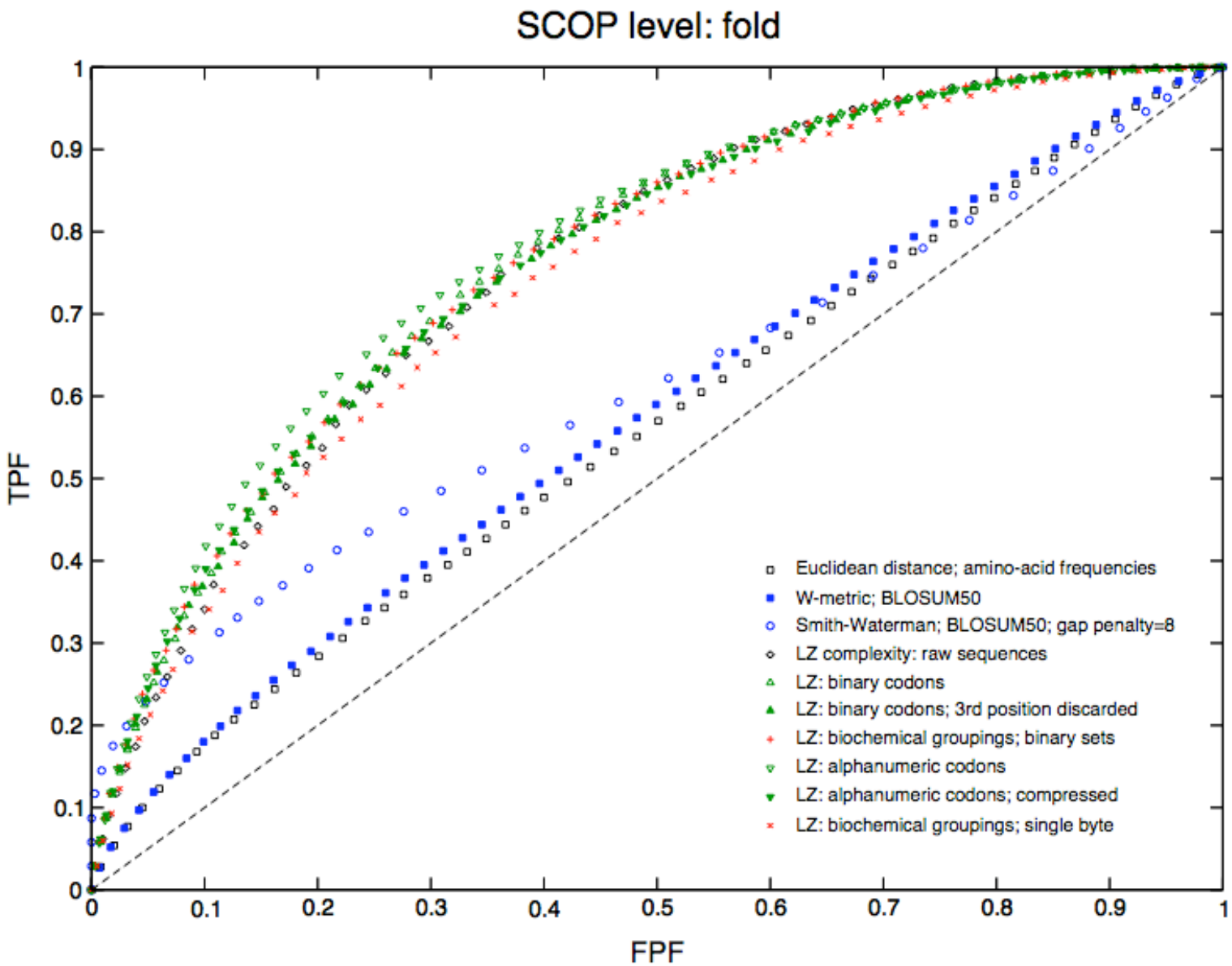
Benchmark graphs



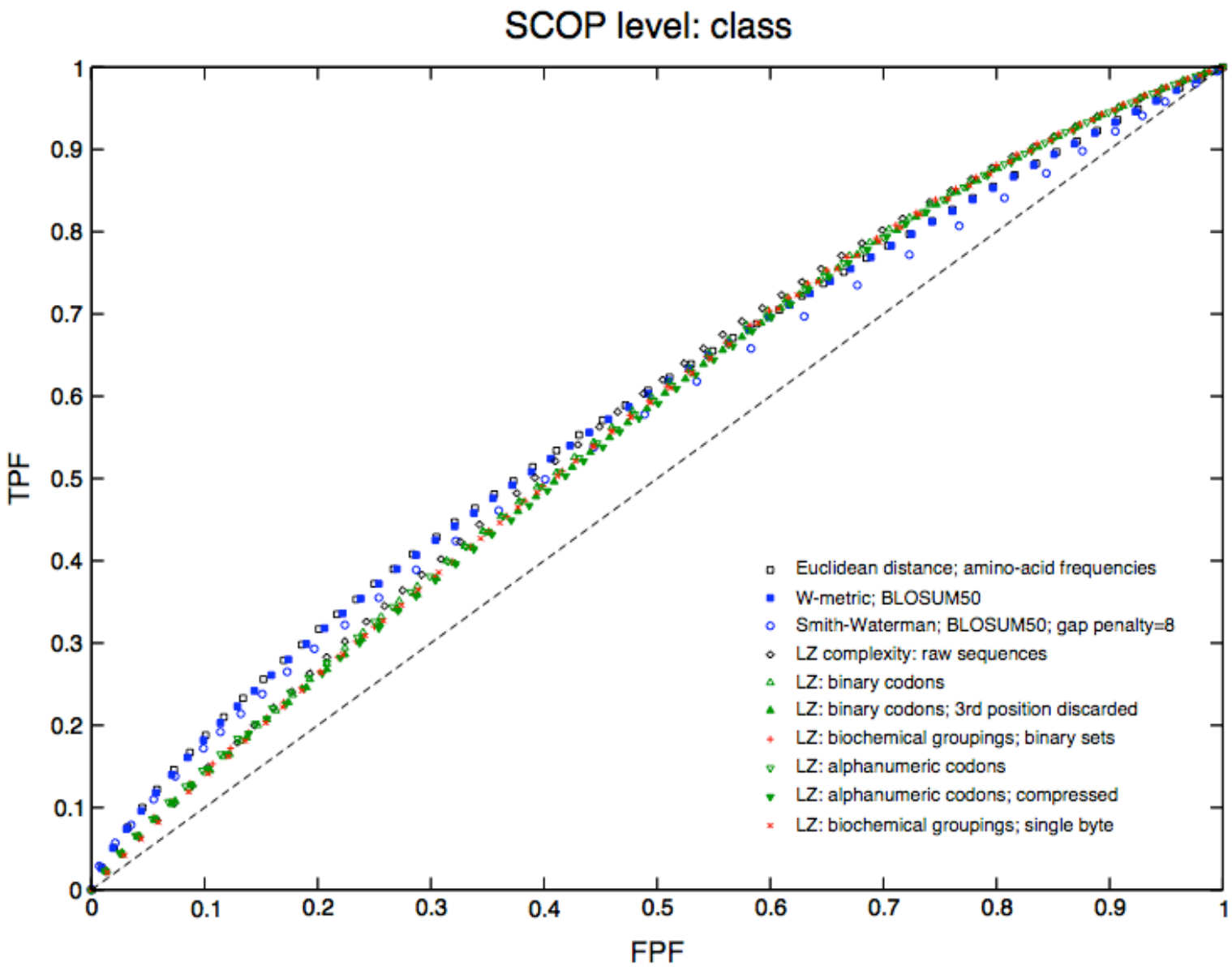
Benchmark graphs



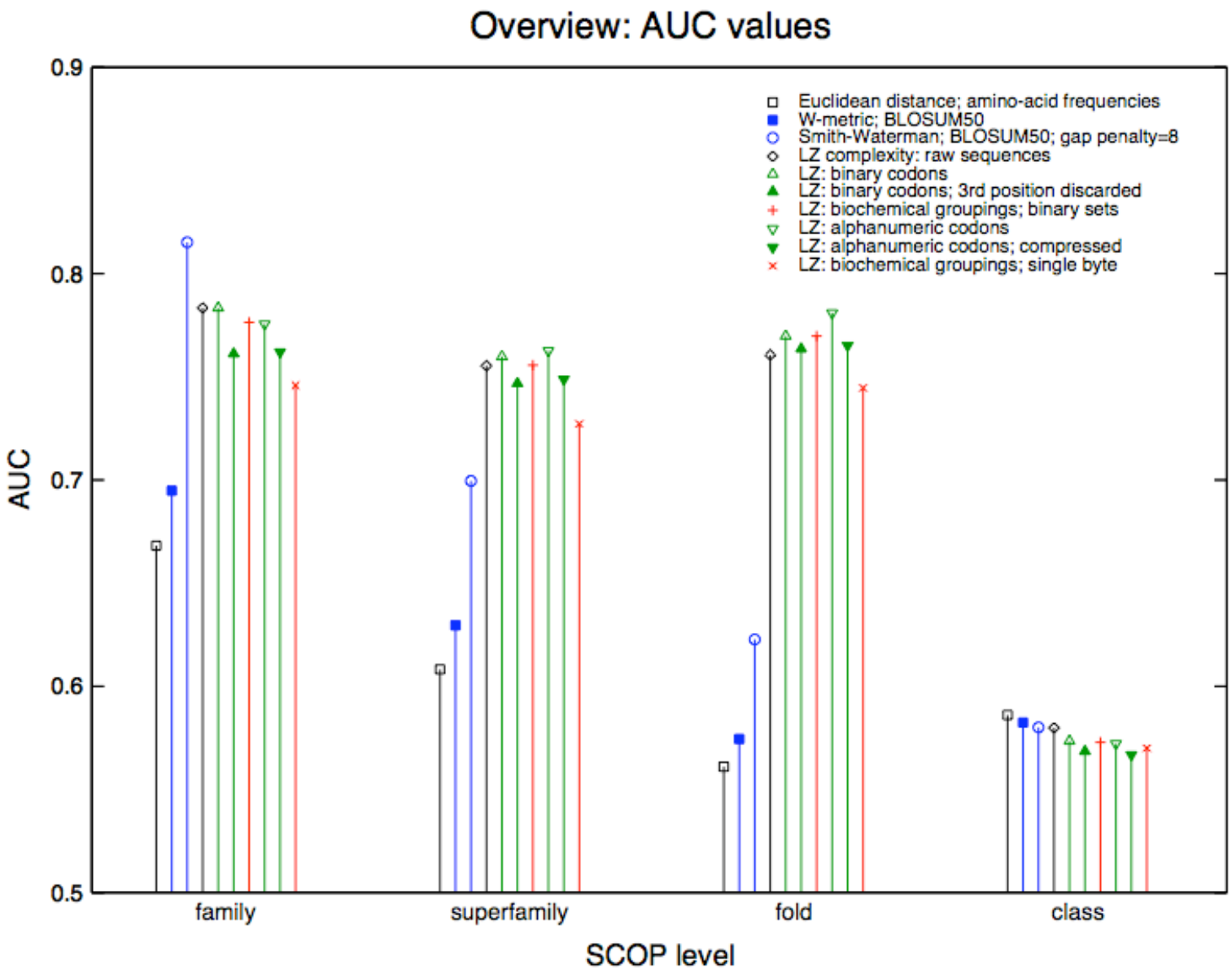
Benchmark graphs



Benchmark graphs

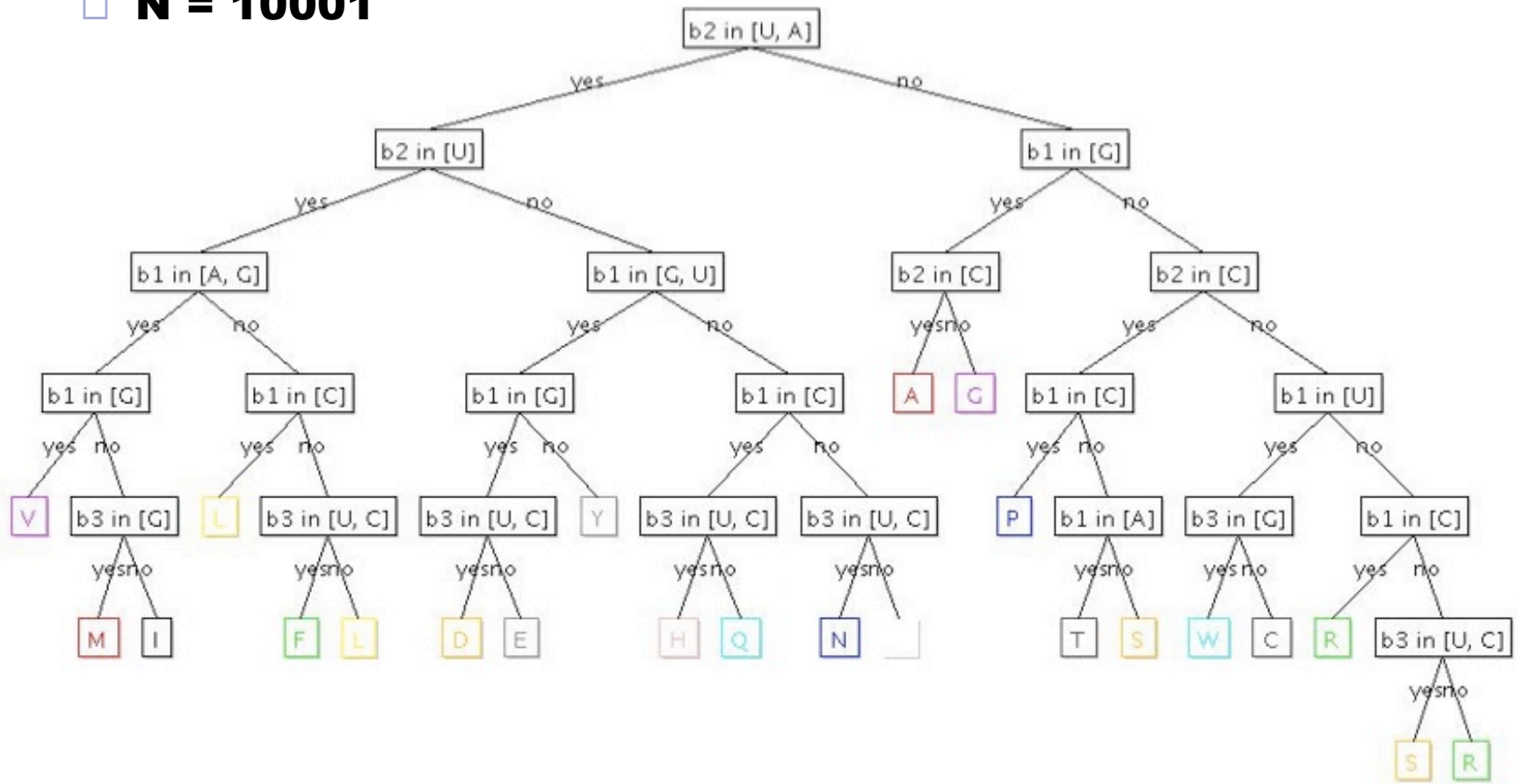


Benchmark graphs

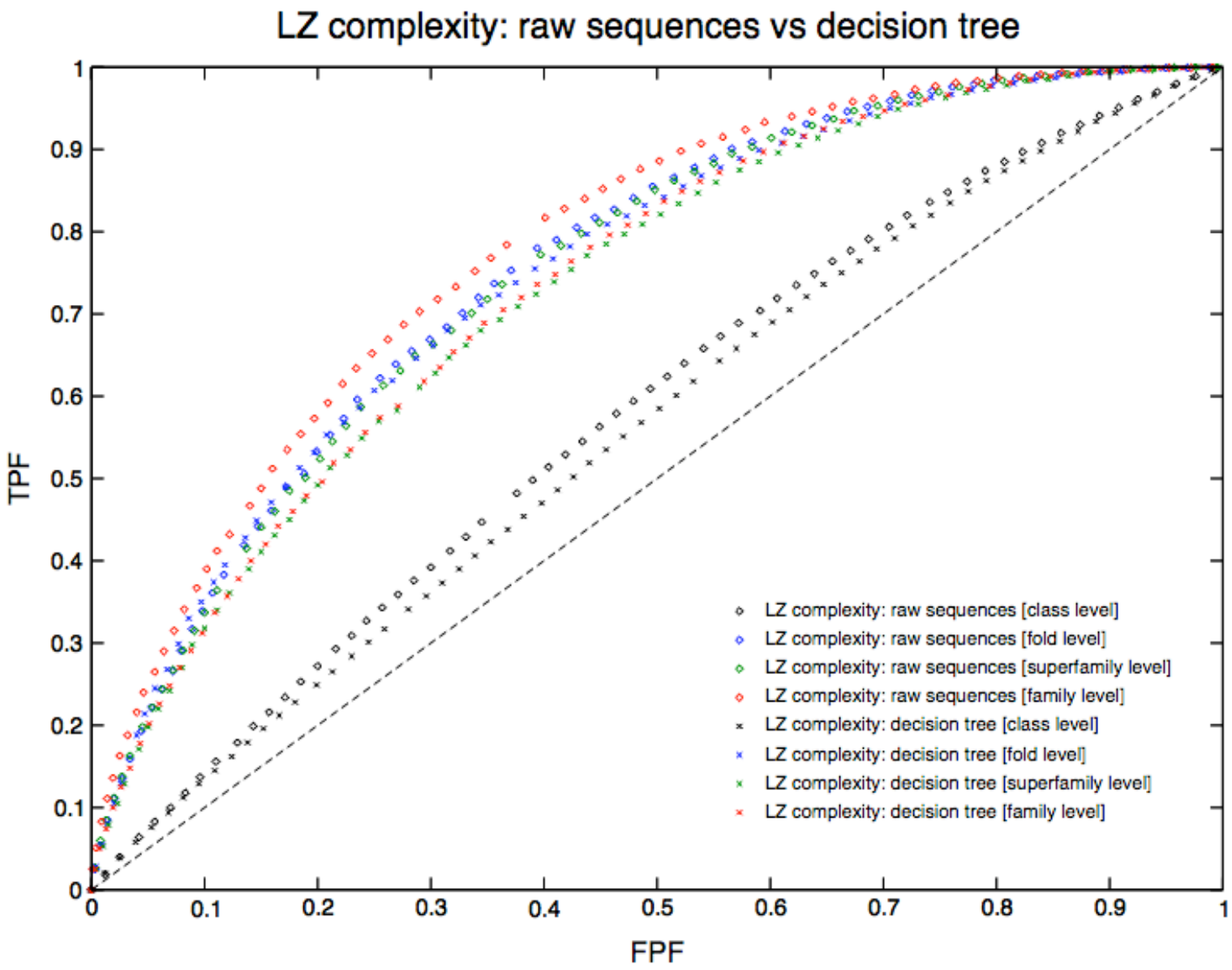


Addendum: decision trees

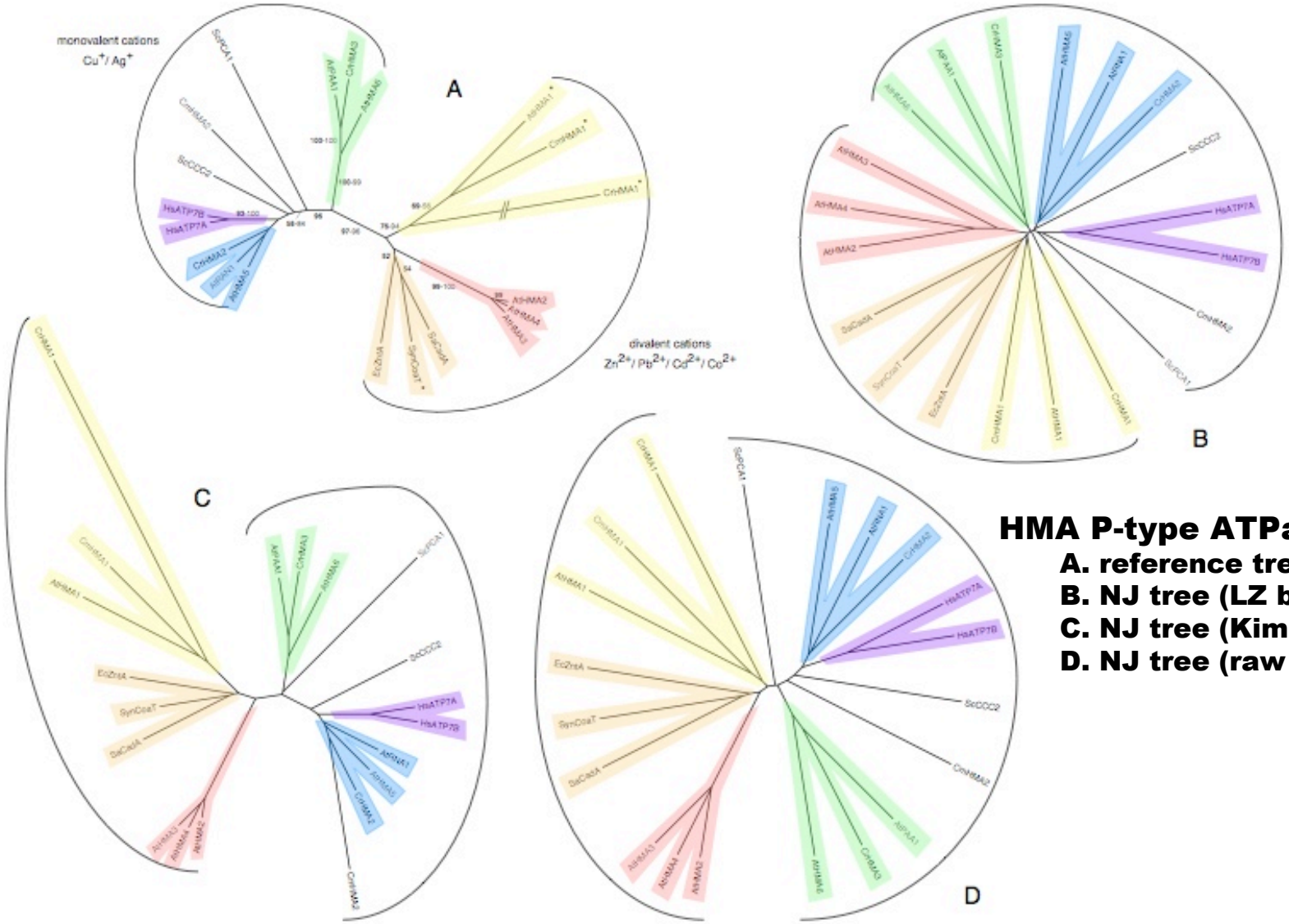
- **codon-based**
  - **R = 00001 (shortest path)**
  - **M = 11101**
  - **N = 10001**



Addendum: decision trees



Application to phylogenetics



HMA P-type ATPases

- **A. reference tree (MP/ML/NJ)**
- **B. NJ tree (LZ binary AA-sets)**
- **C. NJ tree (Kimura distances)**
- **D. NJ tree (raw distances)**

Hanikenne *et al.* (2005) *Plant Physiol.*, **137**, 428-446

Discussion

Performance considerations

	Eu	W_m	SW	LZ	LZ1	LZ2	LZ3	LZ4	LZ5	LZ6	LZ7
CPU-time	0.2	8	264	1	75	37	37	48	16	1	16
language	perl		perl/C	C							

- **one relative unit**
  - **7 min 45 on a PowerPC G4 at 1.25 GHz (Mac OS X)**
  - **6 min 10 on a Pentium 4 at 2.4 GHz (SuSE Linux)**
- **implementation**
  - **perl/c is a perl wrapper for the *water* program (written in C) of the EMBOSS software package**
  - **our software is algorithmically optimized but not technically optimized (further optimizations on the way)**

Conclusions

- **benchmarks**
  - **while computationally affordable, the LZ complexity *outperforms all other methods* at the three lower levels of the SCOP classification, except the very slow SW local alignment at the family level**
  - **at superfamily and fold levels, our sequence encodings show slightly *better results* than the default complexity, but at the expense of considerable *computational burden***
- **phylogenetics**
  - **the LZ complexity is able to retrieve *most clades* found through alignment-based methods but would need some kind of *distance correction* to be really useful**

Perspectives

- **refine the code in order to publish it somewhere as an *Applications Note***
  - **probably not novel enough for a research paper**
- **improvement of the method for its use in phylogenetic inference**
  - **change reconstruction algorithm (Fitch instead of NJ)**
  - **fix suboptimal folding of AA biochemical groups**
    - **e.g. Dayhoff groups (C-ILMV-FWY-AGPST-HKR-DENQ)**
  - **modify the LZ complexity itself to favor large patterns**
    - **what about the effect on distance properties?**
- **other suggestions?**