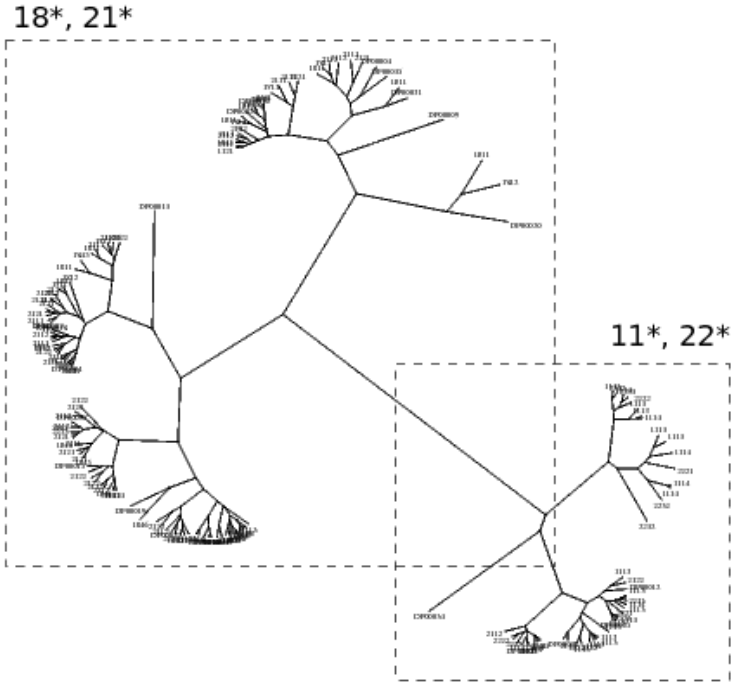
THE SUBSEQUENCE COMPOSITION OF POLYPEPTIDES

FABIO CUNIAL | ALBERTO APOSTOLICO



- Constrained subsequences, positional equivalence, and the ω-suffix graph.
- The dataset and its rationale
- Classifying with suffix graphs
- Laws governing polypeptides
- Laws governing random permutations of polypeptides

2926606920439054397967548375833719159919910473895631392945311389708800463637987473929840325809307030010774766185
391004135689734801234068539673581567775247599312241880503731683150081948935872832347373134284011181695810360683 6

INFORMATION AND PROTEINS: A TORMENTED LOVE AFFAIR

« Perhaps the most remarkable features of the molecule are its complexity and its lack of symmetry. The arrangement seems to be almost totally lacking in the kind of regularities which one instinctively anticipates. »

~ "A three-dimensional model of the myoglobin molecule obtained by X-ray analysis", Kendrew et al., 1958.

« Could the search for ultimate truth really have revealed so hideous and visceral-looking an object? »

~ "The hemoglobin molecule", Perutz, 1964.

« In 1952, F. Sanger described the first complete sequence of a globular protein. This turned out to be both a revelation and a deception. This sequence, defining the structure and therefore the elective properties of a functional protein (insulin), did not show any regularity, characteristic feature, or limit. »

~ "Le hasard et la nécessité", Monod, 1970.

2926606920439054397967548375833719159919910473895631392945311389708800463637987473929840325809307030010774766185
3910041356897348012340685396735815677752475993122418805037316831500819489358728323473731342840111816958103606836

WHAT IS INFORMATION?

- Defining and measuring information in discrete objects is still an open, fuzzy question [Brooks, 2003] [Szpankowski and Konorski, 2007]. Despite our intuition and our hopes.
  - At the two extremes are regular and random strings, to which we would like to give zero complexity, and at the "center" are informative strings [Adami and Cerf, 2000].
- Biological strings are a natural testbed for such measures:
  - obvious applications in classification, prediction, discovery and synthesis;
  - intuitively correct notions, like length of a genome and number of genes, often fail;
  - multiple, concurrent levels of abstraction;
  - the very definition of "random biological string" is elusive;
  - the relationship between information and evolutionary processes, structure, function and chemical activity is a fundamental and widely inexplored domain [Carothers et al. 2004] [Corona et al. 2010, personal communication].

INFORMATION IN POLYPEPTIDES

- Polypeptide strings are particularly challenging:
  - Selection presses polypeptides to streamlined, optimized configurations.
  - Nonlocal interactions.
  - Secondary and tertiary structures occur frequently in random sets of artificial polypeptides.
  - The map from string to function depends on cellular context: chaperones, multimers, post-translational tagging. Anfinsen's dogma [Anfinsen, 1972] is only partially true.
  - Key functions are often implemented by few atoms out of hundreds, in specific spatial configurations and with specific chemical properties ⇒ Sub-symbolic signals.
- Current biochemical consensus: proteins are memorized ancestral random polypeptides, slightly edited by selection to optimize their active sites and stability under physiological conditions [Weiss et al. 2000].

2926606920439054397967548375833719159919910473895631392945311389708800463637987473929840325809307030010774766185
3910041356897348012340685396735815677752475993122418805037316831500819489358728323473731342840111816958103606836

ENTROPY, CORRELATION, PERIODICITY

- Differential entropy and context-free grammar complexity: large sets of nonhomologous proteins are less complex than corresponding sets of random strings by ≈1% [Weiss et al. 2000].
- Ordered proteomes have weak correlations at short, medium (≈100aa) and long range [Weiss et al. 2000].
- Family-dependent, short-range periodicities in hydrophobicity, alpha-helix propensity and charge have been detected in some cases [Weiss and Herzel, 1998] [Rackovsky, 1998], but not in others [White and Jackobs, 1990] [Schwartz and King, 2006].
- Repetitions and redundancies have been implicated in human diseases at the DNA level [Benson and Waterman, 1994] and in the formation of toxic fibrillar structures at the protein level [Broome and Hecht, 2000].
  - They could also hinder the convergence of the folding process into a global minimum [Wan and Wootton, 2000].

COMPRESSIBILITY

Few compressors for polypeptides are available.

- PPM algorithm with contexts of multiple lengths [Nevill-Manning and Witten, 1999].
- Exact/approximate reverse complements and repeats [Matsumoto et al. 2000].
- Partition of aa according to their frequency, plus popular text compressors [Sampath, 2003].
- Huffman codes guided by aa substitution matrices [Hategan and Tabus, 2004].
- Off-line dictionary of extensible subsequences [Apostolico et al. 2006].
- Panels of weigthed experts using species information, local context and repetitions [Cao et al. 2007].
- Applied to proteomes or other large sets: expansions frequently reported in stand-alone strings.
- Entropies achieved (bps): $3.67 \leq H \leq 4.05$. Uniform distribution: $H \approx 4.32$.
- Compression has been shown to grasp structural/functional information via the universal similarity metric [Ferragina et al. 2007].

2926606920439054397967548375833719159919910473895631392945311389708800463637987473929840325809307030010774766185
3910041356897348012340685396735815677752475993122418805037316831500819489358728323473731342840111816958103606836

NUMBER OF SUBSTRUCTURES

- Measuring information/complexity by counting the number of substructures occurring in a string is not a well-explored avenue [Colosimo and De Luca, 2000].
- In this work, we count the number of distinct strings, of any length, that occur as subsequences in a given polypeptide, subject to the following constraints:
  - a maximum gap between consecutive symbols of an ocurrence;
  - greed: the leftmost occurrence of a symbol must be chosen.
- Measures on subsequences are shown:
  - to grasp structural/functional information of polypeptides;
  - to reveal laws followed by structurally and functionally diverse polypeptides;
  - not to separate proteins from their random permutations;
    - but to reveal laws followed by the random permutations of polypeptides.

2926606920439054397967548375833719159919910473895631392945311389708800463637987473929840325809307030010774766185
3910041356897348012340685396735815677752475993122418805037316831500819489358728323473731342840111816958103606836

A string $v \in \Sigma^*$ is an ω-*subsequence* of string s if

$$\exists\, 0 \leq i_0 < i_1 < \ldots < i_{|v|-1} < |s| \;:\; s[i_j] = v[j]\; \forall\, 0 \leq j < |v|$$

and if $i_{j+1} - i_j \leq \omega\; \forall\, 0 \leq j < |v| - 1$.

The *left occurrence list* of an ω-subsequence v is defined as:

$$\mathcal{L}_v := \{i_0 \mid (i_0, i_1, \ldots, i_{|v|-1}) \text{ is an occurrence of } v \text{ in } s \text{ as an } \omega\text{-subsequence}\}\ .$$

The *right occurrence list* $\mathcal{R}_v$ is similarly defined.

The *horizon* $\mathcal{H}_{v,i_0}$ seen by an ω-subsequence v from position i_0 in s, is defined as:

$$\mathcal{H}_{v,i_0} := \{(i_{|v|}, s[i_{|v|}, i_{|v|+\omega-1}]) \mid (i_0, i_1, \ldots, i_{|v|-1}) \text{ is an } \omega\text{-occurrence of } v \text{ in } s\}\ .$$

The *panorama* $\mathcal{P}_v$ seen by an ω-subsequence v in s is similarly defined as:

$$\mathcal{P}_v := \{s[i_{|v|}, i_{|v|+\omega-1}] \mid (i_0, i_1, \ldots, i_{|v|-1}) \text{ is an } \omega\text{-occurrence of } v \text{ in } s\}\ .$$

2926606920439054397967548375833719159919910473895631392945311389708800463637987473929840325809307030010774766185
3910041356897348012340685396735815677752475993122418805037316831500819489358728323473731342840111816958103606836

POSITIONAL EQUIVALENCE, IMPLICATION

Definition 1 (Left equivalence) *Two subsequences v and w are* left equivalent*, denoted $v \equiv_l w$, if $\mathcal{L}_v = \mathcal{L}_w$.*

Property 1 *If $v \equiv_r w$, then $\mathcal{P}_v = \mathcal{P}_w$.*
Property 2 *The relation $\equiv_r$ is right-invariant, i.e., $v \equiv_r w$ implies $va \equiv_r wa \; \forall \; a \in \Sigma$*

Definition 2 (Implication) *We say that w* implicates *or* induces *v on s if for every occurrence $\mathbf{i_1} = \langle i_{11}, i_{12}, ..., i_{1k} \rangle$ of v there is also an occurrence $\mathbf{i_2} = \langle i_{21}, i_{22}, ..., i_{2l} \rangle$ of w such that $(i_{11} = i_{21}) \wedge (i_{1k} = i_{2l})$.*

Definition 3 (Equivalence) *Two subsequences v and w of s are* equivalent*, denoted $v \equiv w$, if they implicate one another.*

Lemma 1 *If $v \equiv w$, then $\mathcal{P}_v = \mathcal{P}_w$, moreover, v and w have the same horizon structure.*

Lemma 2 *The equivalence relation $\equiv$ is right-invariant.*

Note that $v \equiv w \; \Rightarrow (\mathcal{L}_v = \mathcal{L}_w) \wedge (\mathcal{R}_v = \mathcal{R}_w)$, but the converse is not true.

2926606920439054397967548375833719159919910473895631392945311389708800463637987473929840325809307030010774766185391004135689734801234068539673581567775247599312241880503731683150081948935872832347373134284011181695810360683 6

SPECIAL SUBSEQUENCES

Definition 4 (Special subsequence) *A string $v \in \Sigma^*$ occurring in s starting at positions in $\mathcal{L}_v \neq \emptyset$ is a* special *subsequence if $\mathcal{L}_{va} \subset \mathcal{L}_v$ for every symbol $a \in \Sigma$ visible from $\mathcal{P}_v$. String v is a* non-special *subsequence if there is a symbol $a \in \Sigma$ visible from $\mathcal{P}_v$ such that $\mathcal{L}_{va} = \mathcal{L}_v$.*

Property 4 *If av is a special subsequence and $a \in \Sigma$, then the suffix v of av is a special subsequence.*

Property 5 *If v is a special subsequence, then such is also any $w \equiv v$.*

Definition 5 (Antispecial subsequence) *A subsequence v of s is* antispecial *if any extension va of v in s, $a \in \Sigma$ results in $va \equiv_l v$.*

Extensions, prefixes and suffixes of an antispecial subsequence are not necessarily antispecial.

Equivalent subsequences are antispecial, but *right-equivalent* subsequences need not be antispecial.

2926606920439054397967548375833719159919910473895631392945311389708800463637987473929840325809307030010774766185391004135689734801234068539673581567775247599312241880503731683150081948935872832347373134284011181695810360683 6

ORGANIZING SUBSEQUENCES IN A GRAPH

Speciality embodies a criterion to construct all the subsequences and equivalence classes of *s*.

1. Let $i_1, i_2, ..., i_m$ be the occurrences of symbol *a* in *s*.
2. Align the suffixes $s[i_j..|s|\text{-}1] \;\; \forall \; 1 \leq j \leq m$ along the *m* coordinate axes of a multidimensional grid, such that $s[i_j + k]$ occupies position *k* + 1 along coordinate i_j. This space is called ω-suffix space of *s* induced by symbol *a*.
3. Mark a matching point in this space whenever $\mathbf{x} = [x_1, x_2, ..., x_m] \neq \mathbf{0}$ is such that $s[x_i] = a \vee x_i = 0, \; \forall 1 \leq i \leq m$.
4. Define the following partial orders on matching points:

$$\mathbf{x} <_\omega \mathbf{y} \Leftrightarrow (\forall 1 \leq i \leq m)\ (x_i < y_i \leq x_i + \omega) \vee (x_i = y_i = 0) \vee (x_i \neq 0 \wedge y_i = 0)$$

$$\mathbf{x} <_G \mathbf{y} \Leftrightarrow (\mathbf{x} <_\omega \mathbf{y}) \wedge (\nexists\, \mathbf{z} \mid \mathbf{x} <_\omega \mathbf{z} <_\omega \mathbf{y} \;\wedge\; \text{label}(\mathbf{z}) = \text{label}(\mathbf{y}))$$

$$\mathbf{x} <_H \mathbf{y} \Leftrightarrow (\mathbf{x} <_G \mathbf{y}) \wedge (\nexists\, \mathbf{z} \mid \mathbf{x} <_G \mathbf{z} <_G \mathbf{y})$$

Each partial order $<_i$ corresponds to a DAG embedded into the suffix space.

We call this DAG ω-suffix graph induced by the triplet (*s*, *a*, $<_i$).

RELATIONSHIPS BETWEEN GRAPH AND SUBSEQUENCES

For subsequences of *s* starting with symbol *a*:

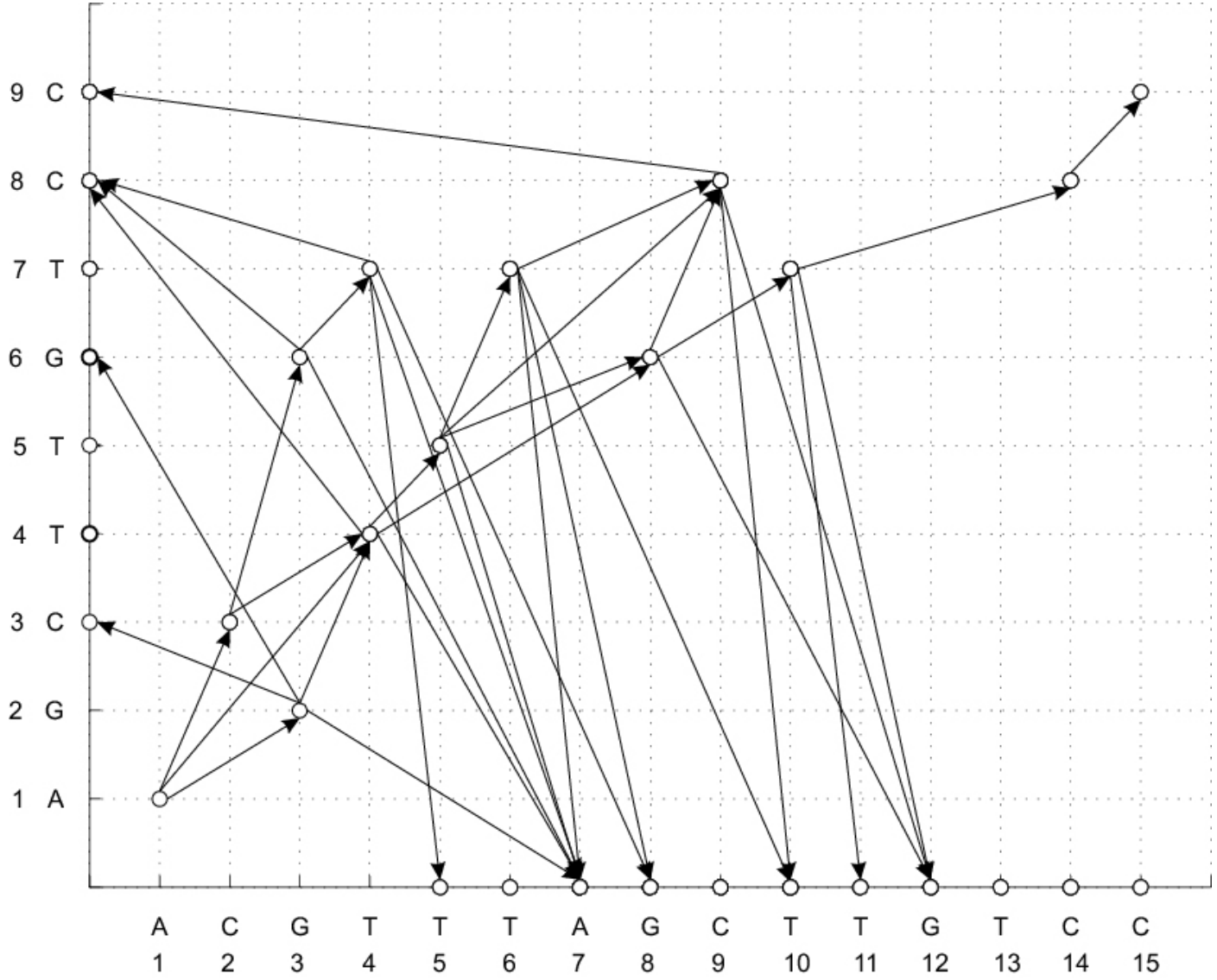
- ω-subsequence of *s* ≡ Chain in $<_\omega$ starting at the origin.
- Greedy w-subsequence of *s* ≡ Chain in $<_G$ starting at the origin.
- Biequivalence class ≡ Matching point reachable from the origin via $<_\omega$.
- Left-equivalence class ≡ Subspace.

Properties:

- Sparsity: in every subspace with *k* dimensions, only matching points in the hyperpyramid with apex at the origin and edges defined by the *k* lines passing through the apex and the points [w,2,2,...,2], [2,w,2,...,2], ..., [2,2,...,w] are visited.
- Redundancy: $\mathcal{O}(\omega|s|^3)$ points suffice to reconstruct any suffix graph without seeing the symbols on the axes.
- At ω=1, a suffix graph collapses into the digital search tree of the substrings of *s*.
- The set of paths of $<_G$ ending in the same point corresponds to a prefix-free set of subsequences.

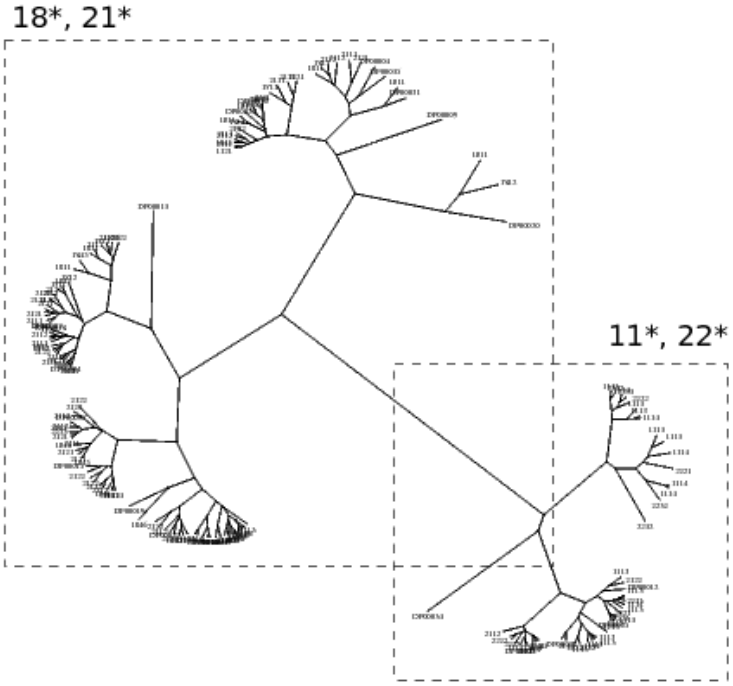
2926606920439054397967548375833719159919910473895631392945311389708800463637987473929840325809307030010774766185
3910041356897348012340685396735815677752475993122418805037316831500819489358728323473731342840111816958103606836

EXAMPLE: GREEDY 4-SUFFIX GRAPH



$s =$ ACGTTTAGCTTGTCC

THE DATASET AND ITS RATIONALE



- Objectives
- Domains: SCOP.
- Disordered regions: DisProt.

2926606920439054397967548375833719159919910473895631392945311389708800463637987473929840325809307030010774766185
3910041356897348012340685396735815677752475993122418805037316831500819489358728323473731342840111816958103606836

OBJECTIVES

Test whether measures on suffix graphs:

- grasp structural/functional features of polypeptides;
- separate polypeptides from their random permutations.

Identify laws governing suffix graph measures in polypeptides.

2926606920439054397967548375833719159919910473895631392945311389708800463637987473929840325809307030010774766185
3910041356897348012340685396735815677752475993122418805037316831500819489358728323473731342840111816958103606836

MEASURES ON SUFFIX GRAPHS

We quantify the compositional richness of a suffix graph, at different values of ω, by counting:

- the number of points: special, antispecial, normal, terminal.
- the number of arcs: internal, external.
- the number of subsequences (paths): special, antispecial, normal, terminal.

We choose not to recode the alphabet with physico-chemical scales [Li et al. 2003].

- Too many scales [Kawashima et al. 2008] and parameters.
- We are after parameter-free, purely syntactic measures.

2926606920439054397967548375833719159919910473895631392945311389708800463637987473929840325809307030010774766185
3910041356897348012340685396735815677752475993122418805037316831500819489358728323473731342840111816958103606836

DOMAINS

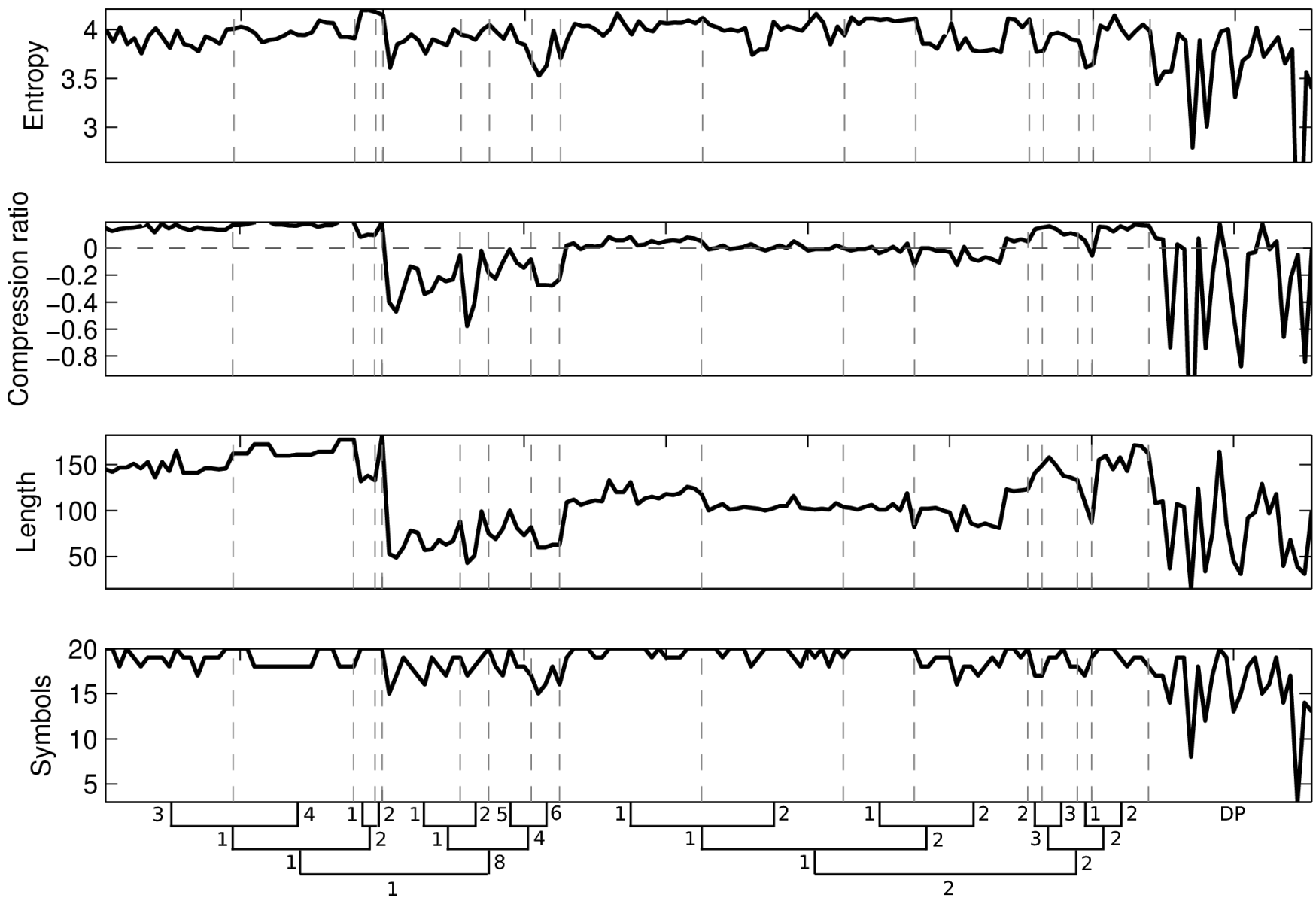
- Domains are recurrent modular subunits of proteins with believed autonomous spatial conformation and function [Richardson, 1981].
- The SCOP database [Murzin et al. 1995] organizes domains in a tree whose levels reflect evolutionary, functional and structural similarities.
  - SCOP is manually curated by biologists.
  - Only the last level contains highly similar strings.
- We analyze an arbitrary subset of 148 domains of SCOP (called D_1 in what follows), spanning different families, superfamilies, folds and classes, to establish:
  - the measures that reconstructs the SCOP tree better;
  - the level of the tree that is reconstructed more accurately;
  - the portions of the SCOP tree that are better separated from their random permutations.

DISORDERED REGIONS

- Proteins contain also “disordered” regions, with no fixed spatial configuration under physiological conditions [Wright and Dyson, 1999].
  - Capable of dynamically transitioning through an ensemble of structures: e.g. they can fold and bind to a target simultaneously.
  - Different sequence-structure mapping that domains: typically enriched in charged and polar, depleted in hydrophobic residues [Li et al. 2000].
  - Tend to have low entropy [Weathers et al. 2006], but there are exceptions.
- DisProt [Sickmeier et al. 2007] is a comprehensive, manually curated, functional classification of all polypeptide regions for which there is experimental evidence of disorder.
- We analyze an arbitrary subset of 23 strings from DisProt (called D_2 in what follows), to establish:
  - the measure on suffix graphs that better separate disordered regions from domains;
  - whether and how disordered regions can be distinguished from random strings.

2926606920439054397967548375833719159919910473895631392945311389708800463637987473929840325809307030010774766185
3910041356897348012340685396735815677752475993122418805037316831500819489358728323473731342840111816958103606836

OVERVIEW OF THE DATASET

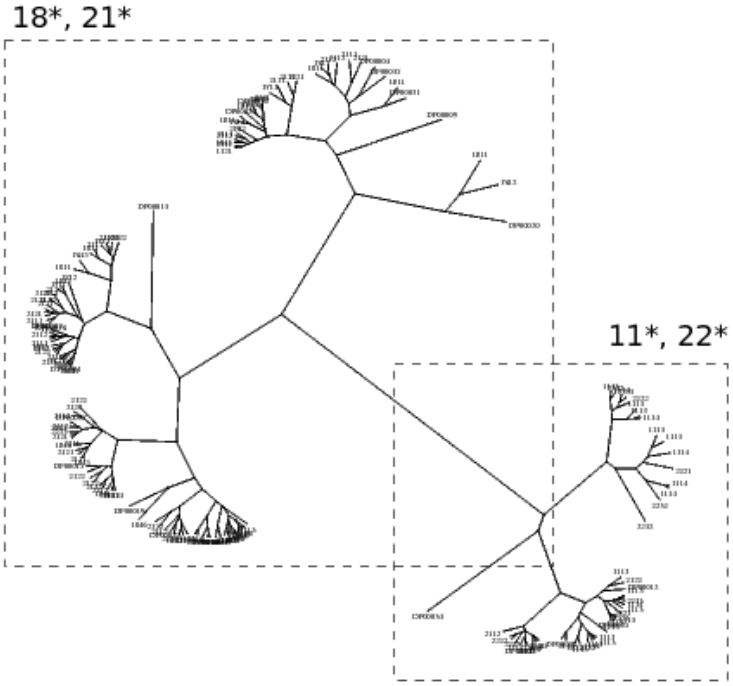


A NOTE ON SIZE

- The analyses that follow aggregate and systematize over one million data points.
- This is perhaps the first time in which the vocabulary of all distinct subsequences of a set of structurally and functionally diverse polypeptides is systematically counted and compared.

2926606920439054397967548375833719159919910473895631392945311389708800463637987473929840325809307030010774766185
3910041356897348012340685396735815677752475993122418805037316831500819489358728323473731342840111816958103606836

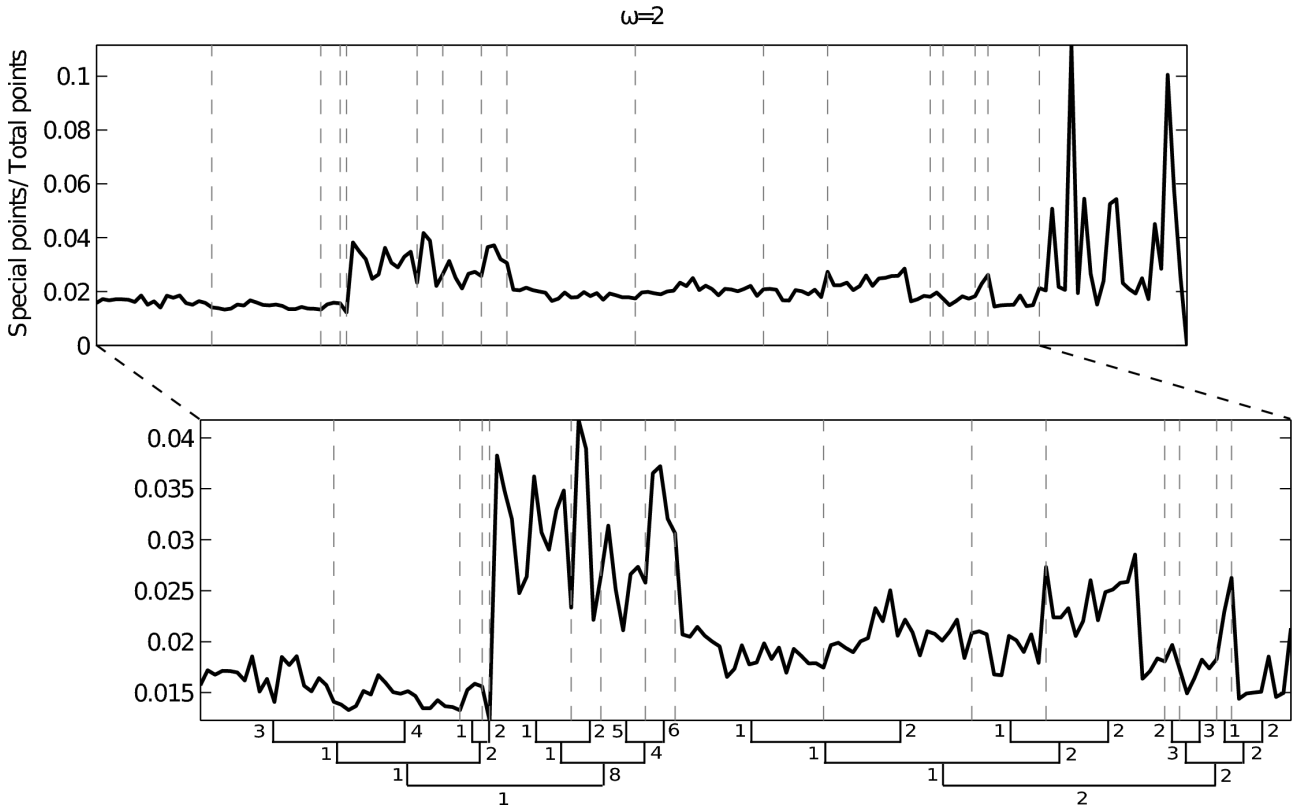
CLASSIFYING WITH SUFFIX GRAPHS



- Stand-alone raw measures
- Abundance vectors
- Relative number of points
- Relative number of subsequences
- Transition at ω=5
- Comparison to existing distance measures

STAND-ALONE RAW MEASURES

- Stand-alone raw measures seem unrelated to SCOP: no clustering.
- Some stand-alone measures normalized by the total number of their respective elements seem to show a clustered behavior.
- Clusters are largely consistent across different measures.



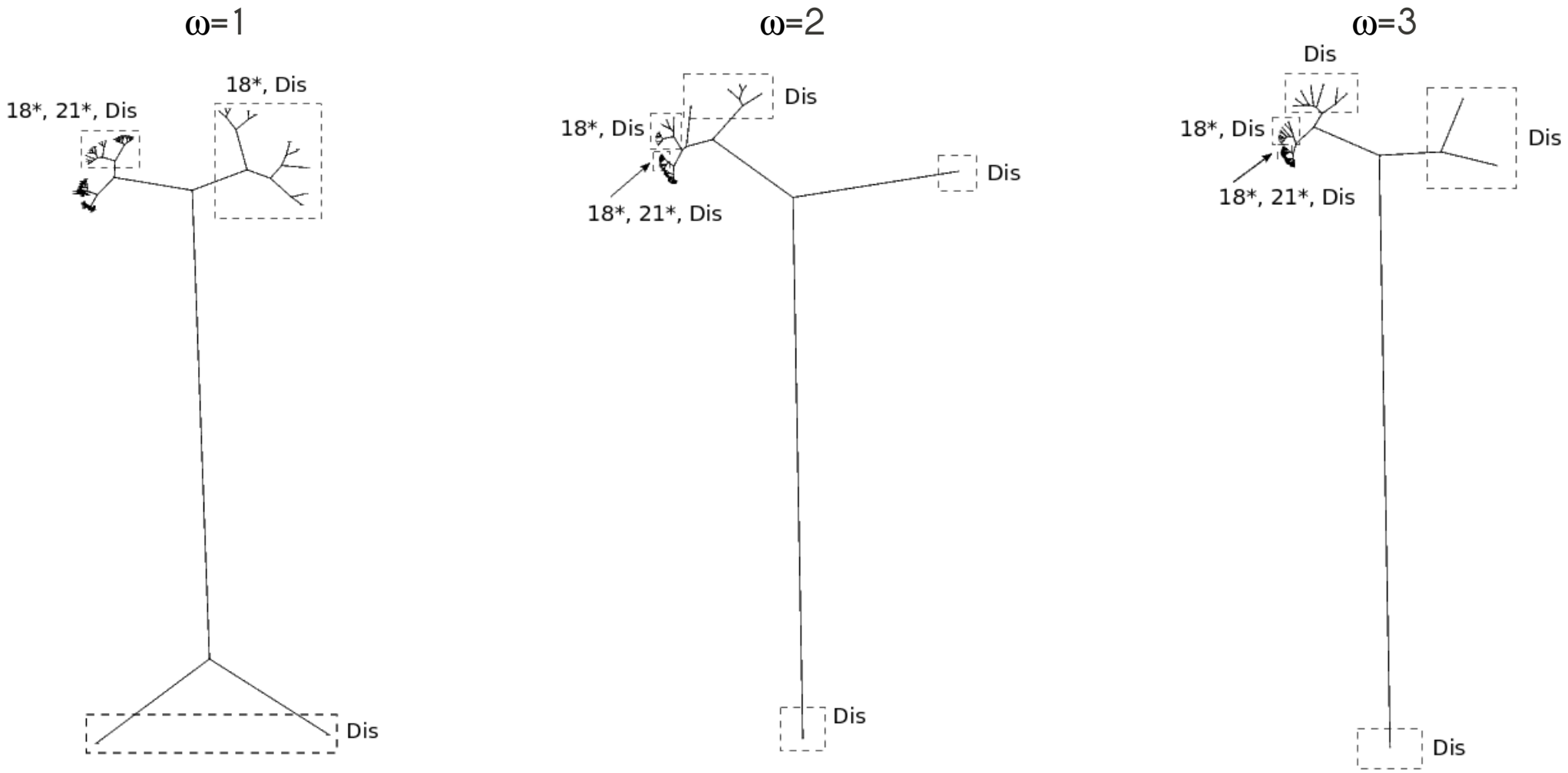
2926606920439054397967548375833719159919910473895631392945311389708800463637987473929840325809307030010774766185
3910041356897348012340685396735815677752475993122418805037316831500819489358728323473731342840111816958103606836

QUANTIFYING THE CLUSTERS: ABUNDANCE VECTORS

- To quantify these clusters, we project each string into an “abundance vector” in the simplex of the relative number of special, antispecial, normal, terminal points.
- We build UPGMA trees using Euclidean distance between abundance vectors.

2926606920439054397967548375833719159919910473895631392945311389708800463637987473929840325809307030010774766185
3910041356897348012340685396735815677752475993122418805037316831500819489358728323473731342840111816958103606836

RELATIVE NUMBER OF POINTS



Strong outliers: { gggsgggsgggsegggsegggsegggsegggsgggsgsg, dprfqdsssskapppslpspsrlpgpsdtpilpq, ... }

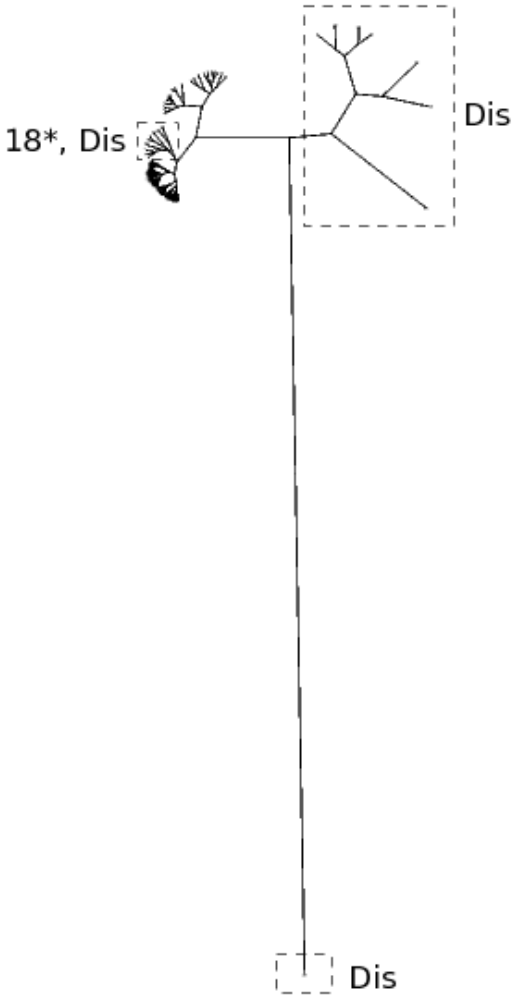
Strong outliers have more special, normal and terminal points.

Fold 1.8 separated from the rest of D_1: more terminal points.

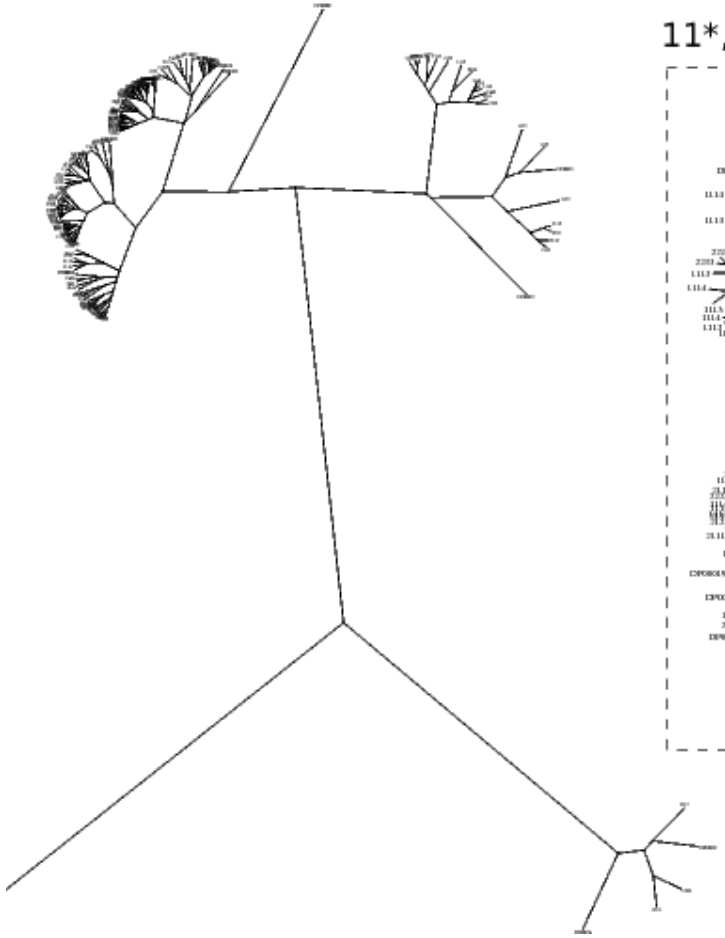
2926606920439054397967548375833719159919910473895631392945311389708800463637987473929840325809307030010774766185
3910041356897348012340685396735815677752475993122418805037316831500819489358728323473731342840111816958103606836

RELATIVE NUMBER OF POINTS

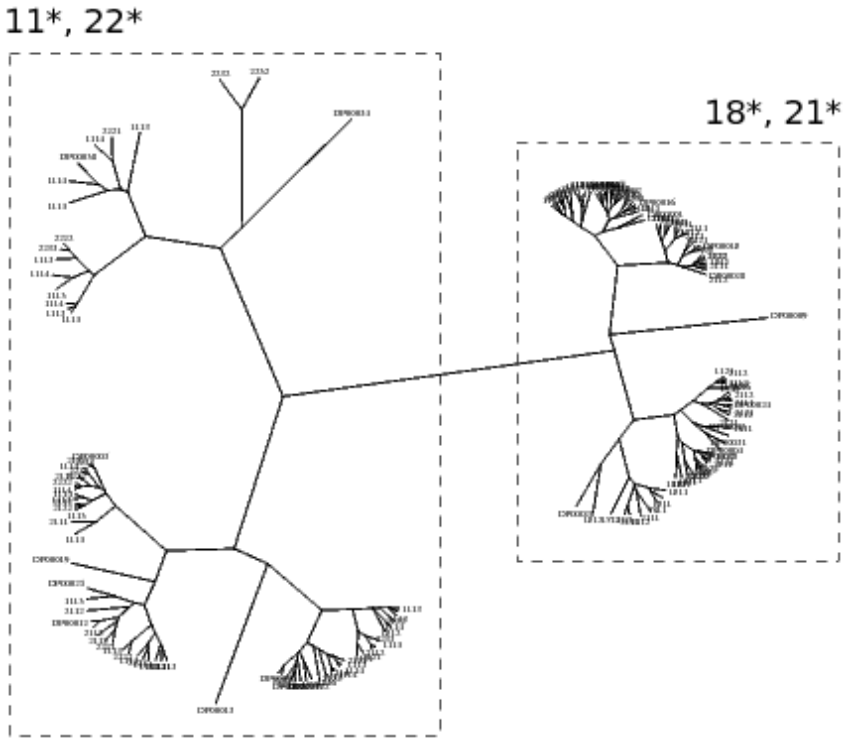
ω=4



ω=5



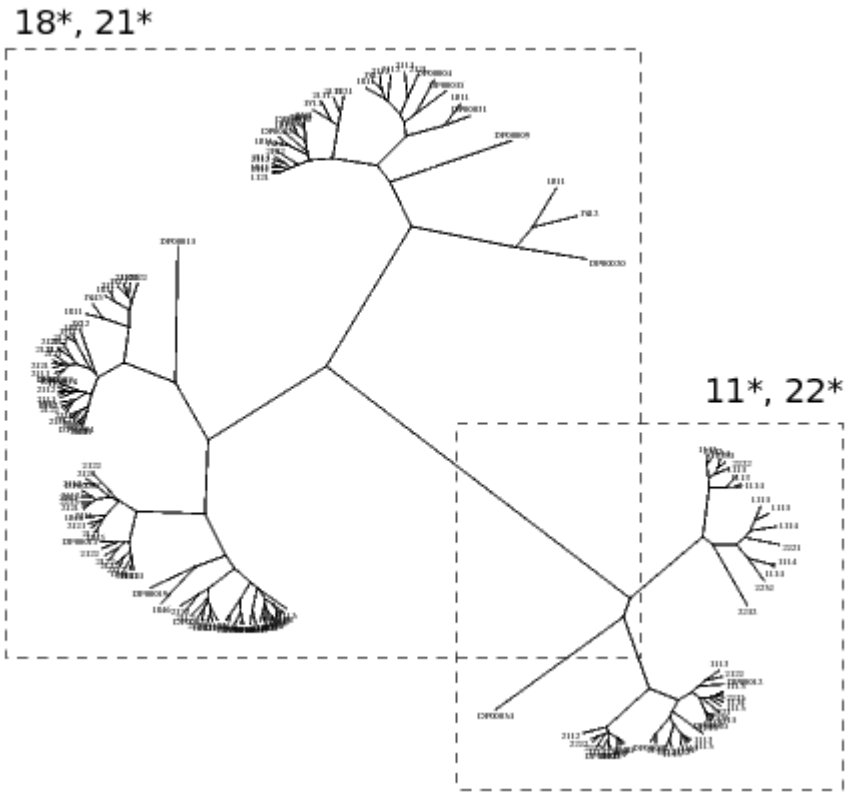
ω=6



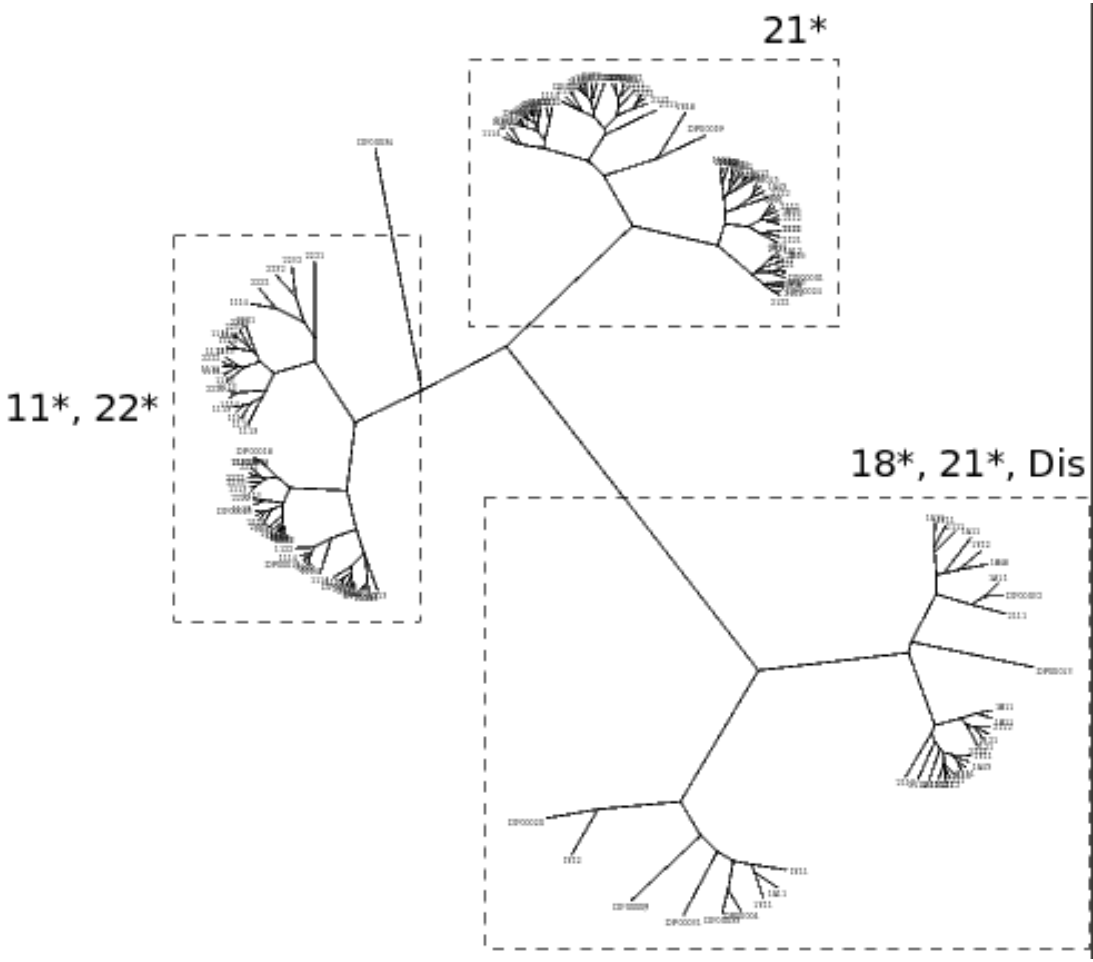
2926606920439054397967548375833719159919910473895631392945311389708800463637987473929840325809307030010774766185
3910041356897348012340685396735815677752475993122418805037316831500819489358728323473731342840111816958103606836

RELATIVE NUMBER OF POINTS

ω=7

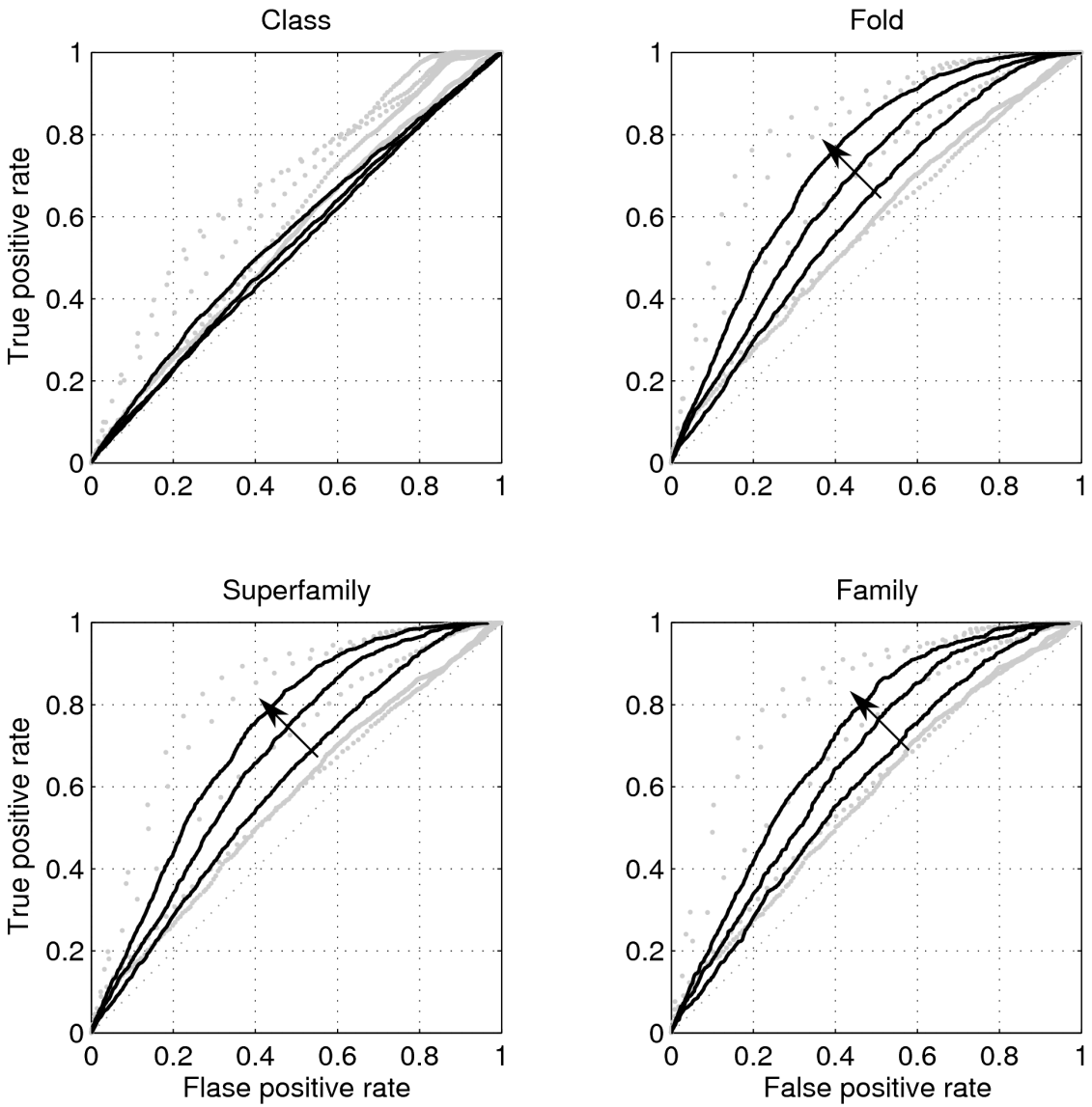


ω=8



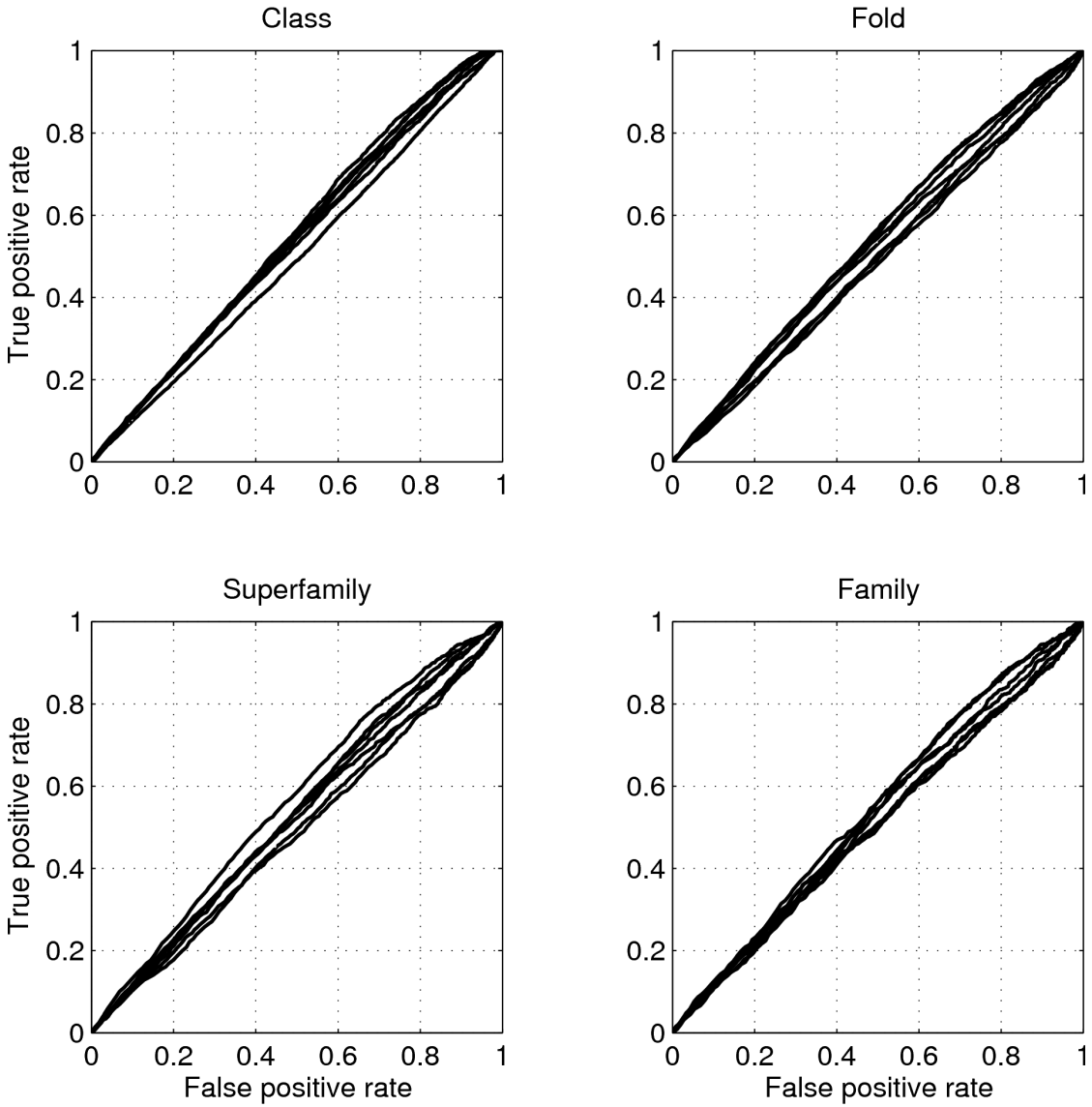
2926606920439054397967548375833719159919910473895631392945311389708800463637987473929840325809307030010774766185
3910041356897348012340685396735815677752475993122418805037316831500819489358728323473731342840111816958103606836

QUANTIFYING THE TRANSITION AT ω=5



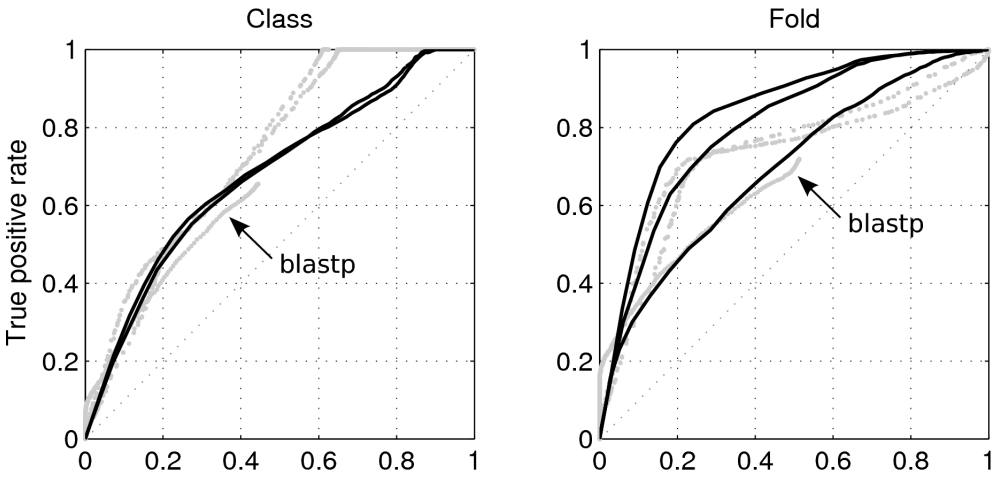
[Ferragina et al. 2007]

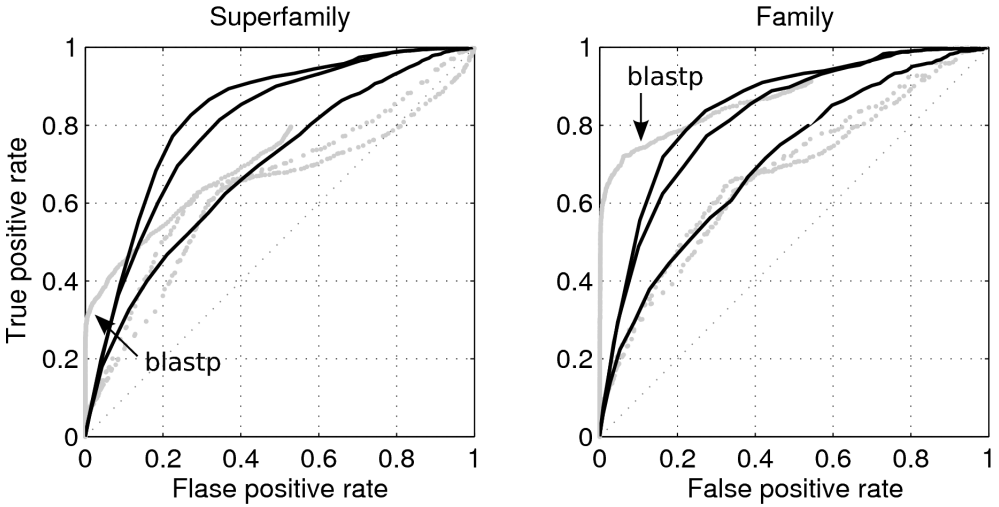
TREES BUILT WITH SUBSEQUENCES DO NOT RESEMBLE SCOP



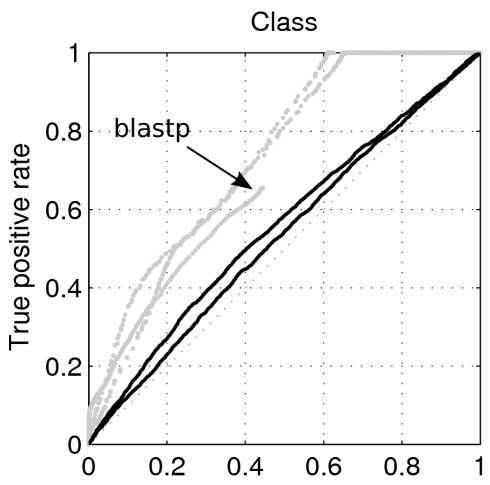
2926606920439054397967548375833719159919910473895631392945311389708800463637987473929840325809307030010774766185
3910041356897348012340685396735815677752475993122418805037316831500819489358728323473731342840111816958103606836

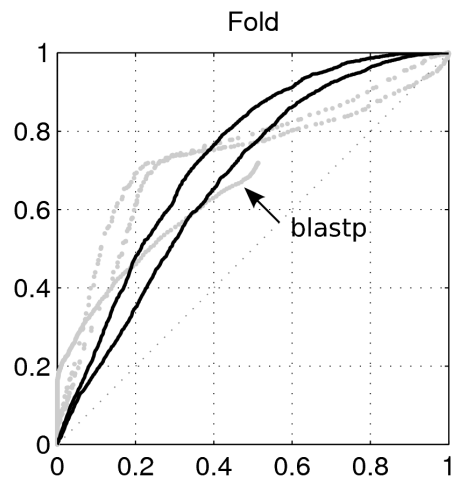
EXISTING DISTANCE MEASURES

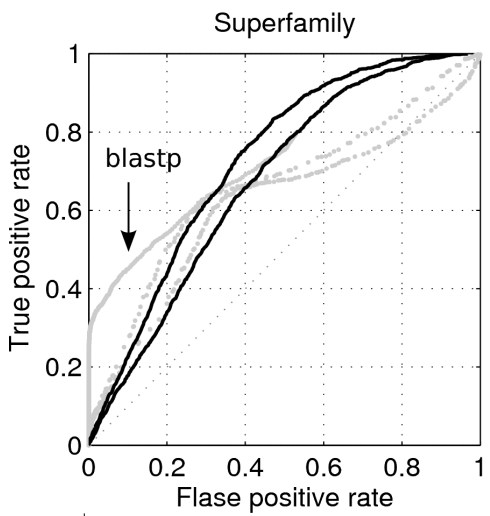


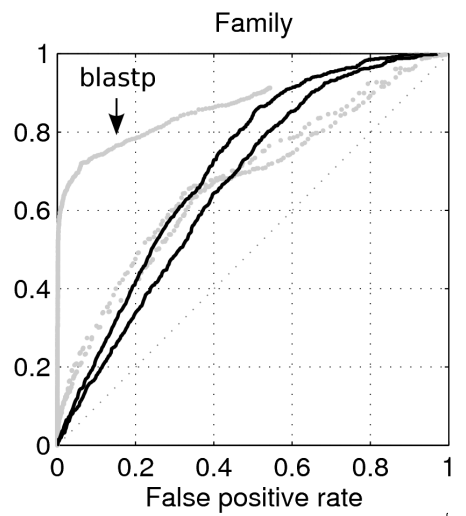


ω <= 3









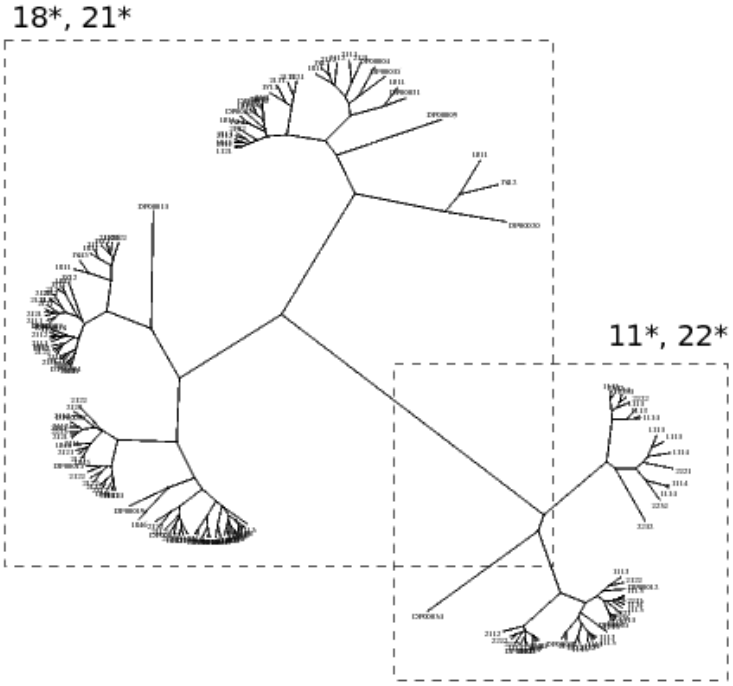
ω >= 7

2926606920439054397967548375833719159919910473895631392945311389708800463637987473929840325809307030010774766185
3910041356897348012340685396735815677752475993122418805037316831500819489358728323473731342840111816958103606836

SURPRISES

- Measures of compositional complexity do grasp SCOP information.
- Contrary to blastp and NCD [Li et al. 2003], they don't use patterns or substructures common to two strings.
- Purely syntactic measures: no recoding of aa with chemical scales.
- Strong outliers.
- Transition in the shape of the trees.
- Difference between points and subsequences.

LAWS GOVERNING POLYPEPTIDES



- Dependence on string length
- Single locus for diverse polypeptides
- Outliers: loci are not unavoidable.
- Shape of the loci depends on ω
- Transitions of abundances in the same string
- Single mesures vs ω in the whole dataset

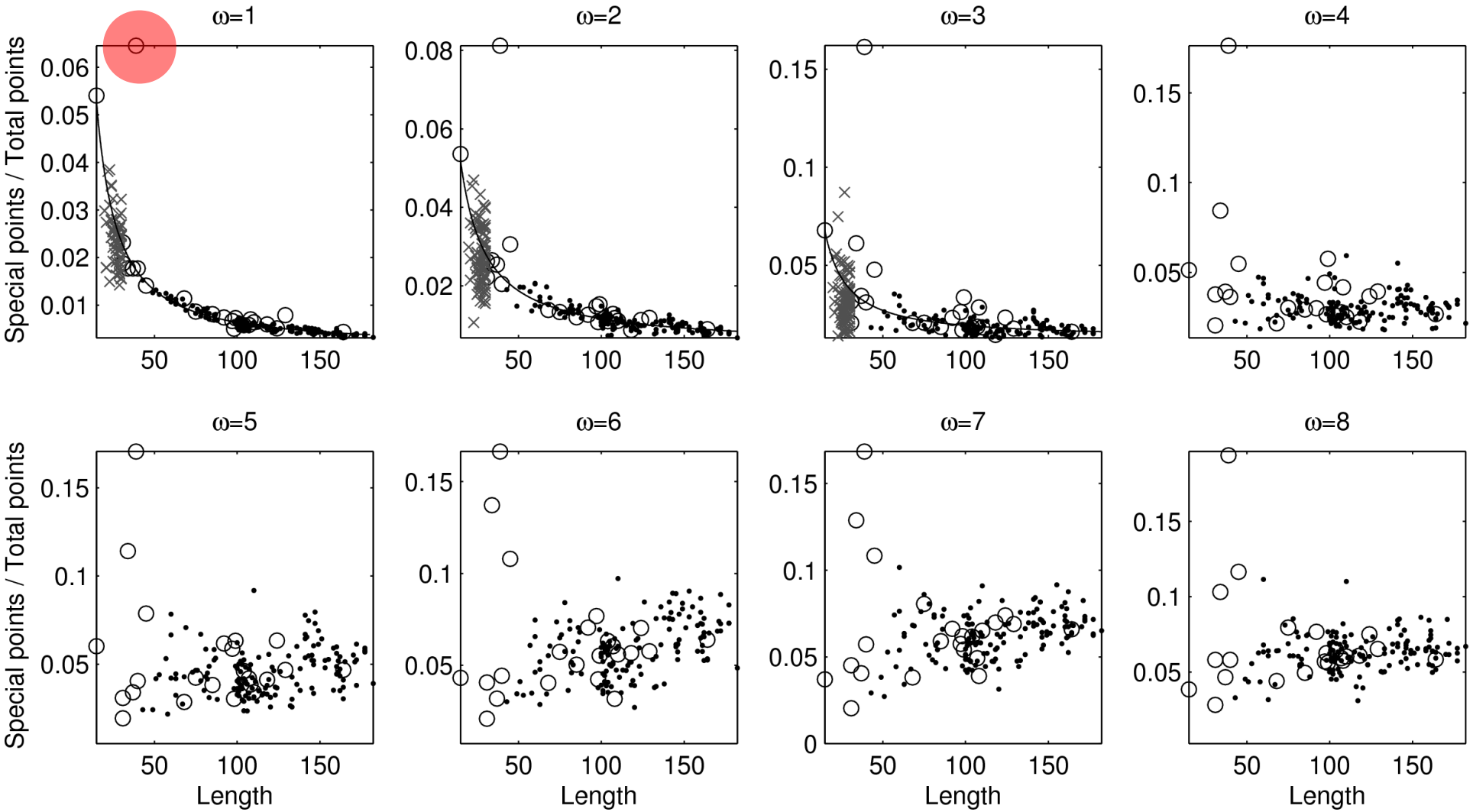
2926606920439054397967548375833719159919910473895631392945311389708800463637987473929840325809307030010774766185
3910041356897348012340685396735815677752475993122418805037316831500819489358728323473731342840111816958103606836

LAWS GOVERNING POLYPEPTIDES

- Plots of normalized stand-alone measures suggest direct and inverse proportionalities with respect to string length.
- E.g. the relative number of special points seems inversely proportional to string length.
  - Expected at $\omega=1$, with equation $y=a/n+b$.
  - However, there could be a different pair (a,b) for each polypeptide, forming a cloud in the (y,n) space.
- Are there common curves ("loci") in the (y,n) space?
- How do these loci behave when ω changes?

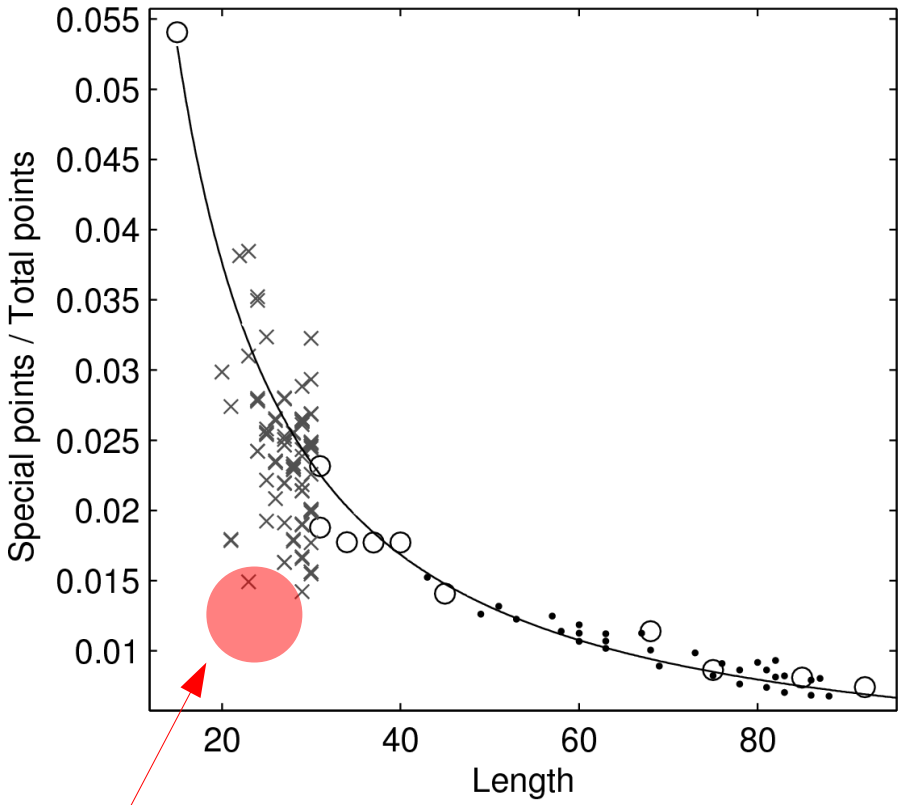
2926606920439054397967548375833719159919910473895631392945311389708800463637987473929840325809307030010774766185
3910041356897348012340685396735815677752475993122418805037316831500819489358728323473731342840111816958103606836

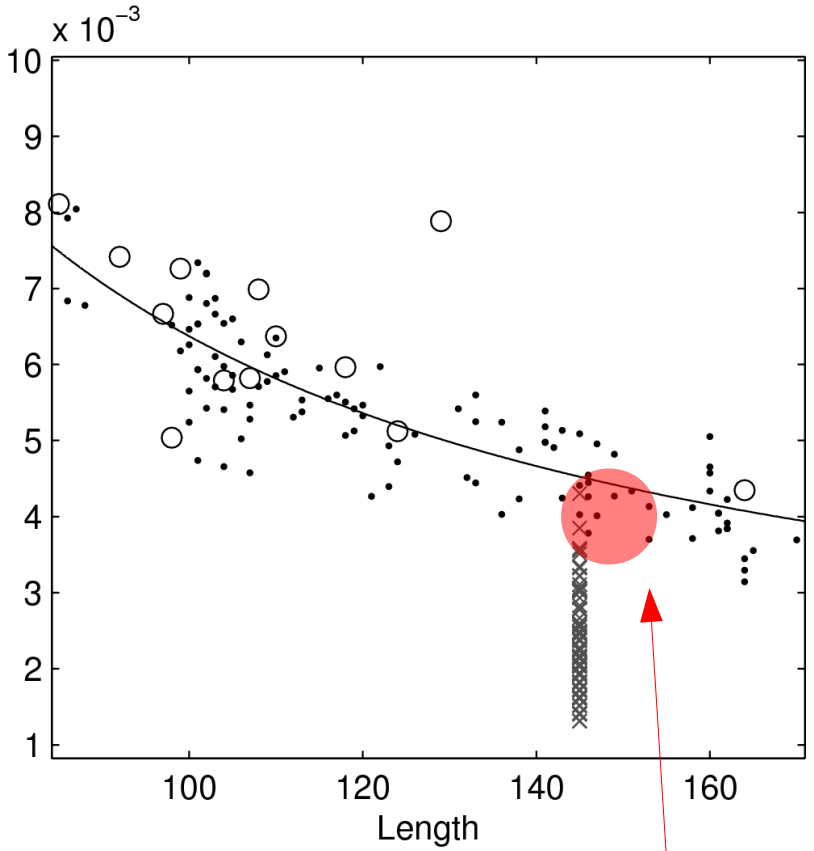
(SPECIAL POINTS / TOTAL POINTS) vs LENGTH



2926606920439054397967548375833719159919910473895631392945311389708800463637987473929840325809307030010774766185
3910041356897348012340685396735815677752475993122418805037316831500819489358728323473731342840111816958103606836

(SPECIAL POINTS / TOTAL POINTS) vs LENGTH, w=1



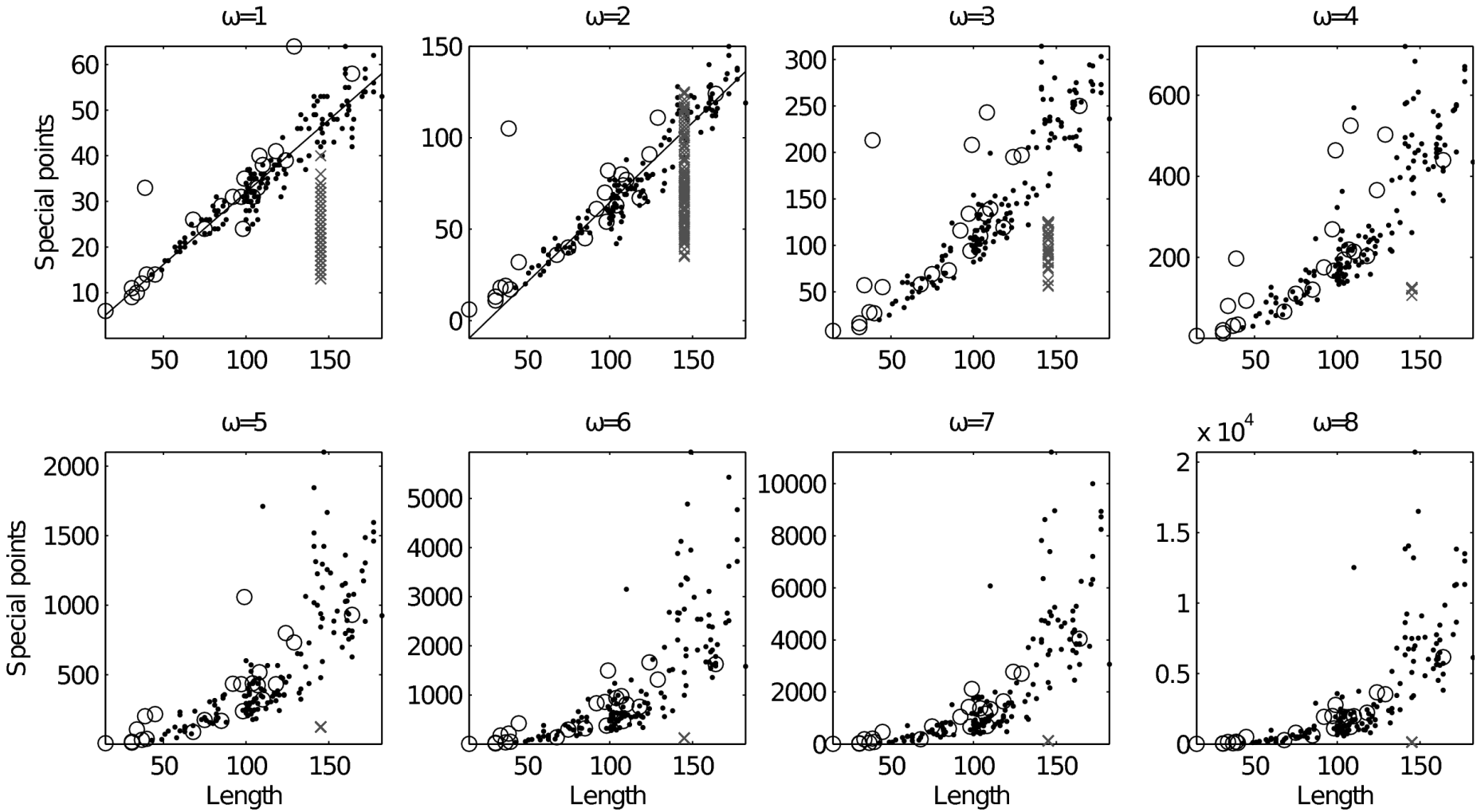


- No strong periodic structure.
- Same number of symbols as D_1 U D_2.
- Same entropy as D_1 U D_2.

Random strings on 20 symbols and minimum entropy.

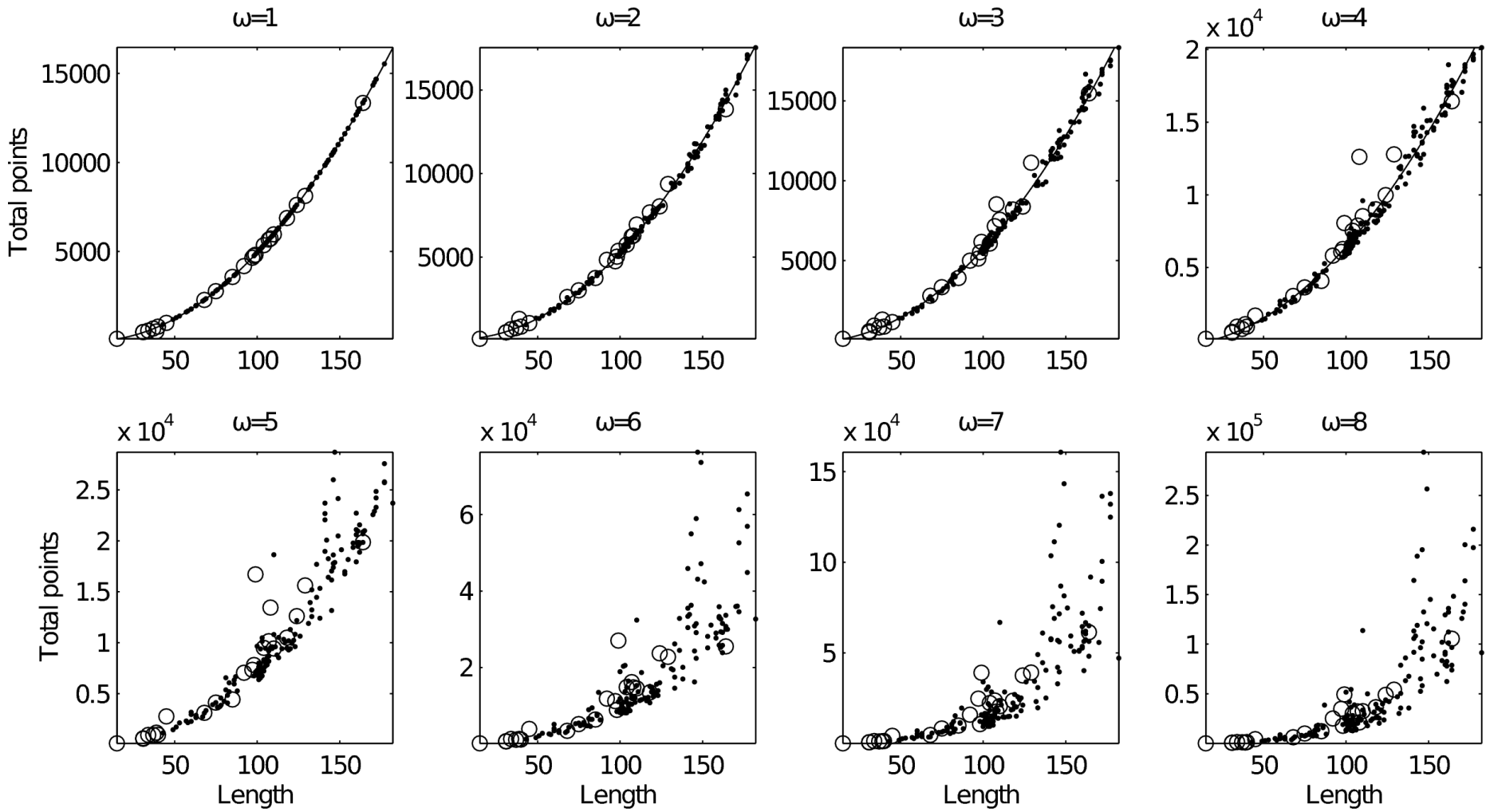
2926606920439054397967548375833719159919910473895631392945311389708800463637987473929840325809307030010774766185
3910041356897348012340685396735815677752475993122418805037316831500819489358728323473731342840111816958103606836

SPECIAL POINTS vs LENGTH



2926606920439054397967548375833719159919910473895631392945311389708800463637987473929840325809307030010774766185
3910041356897348012340685396735815677752475993122418805037316831500819489358728323473731342840111816958103606836

TOTAL POINTS vs LENGTH



2926606920439054397967548375833719159919910473895631392945311389708800463637987473929840325809307030010774766185
391004135689734801234068539673581567775247599312241880503731683150081948935872832347373134284011181695810360683 6

DEPENDENCE ON LENGTH EXPLAINS TREE TRANSITION

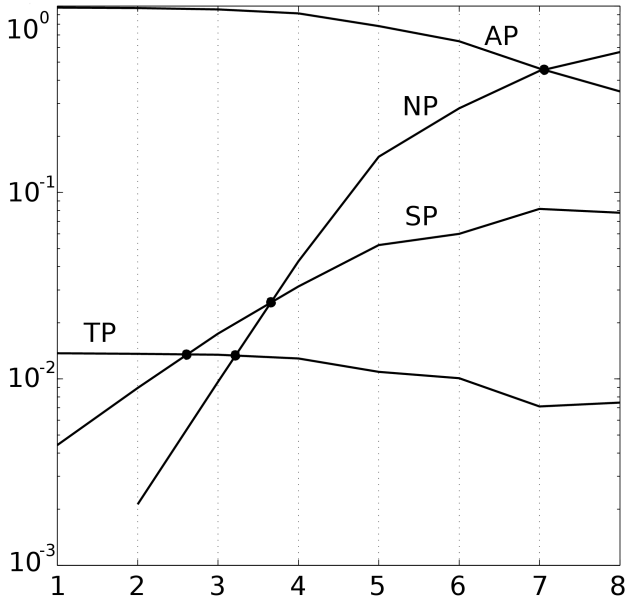
- Similar behaviors appear when other measures on points are considered.
- These dependencies on string length explain the transition seen in classification trees.
  - At $\omega<=3$ special and antispecial points are most numerous, and their abundance is strictly controlled by string length.
  - At $\omega>=7$ antispecial and normal points are most numerous, and their abundance is loosely controlled by string length.
  - At $4<=\omega<=6$ no abundance is strictly controlled by string length.

2926606920439054397967548375833719159919910473895631392945311389708800463637987473929840325809307030010774766185
3910041356897348012340685396735815677752475993122418805037316831500819489358728323473731342840111816958103606836

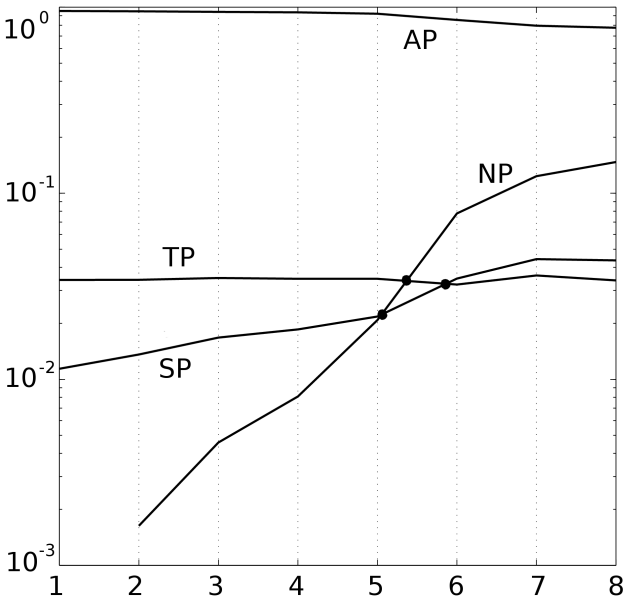
DEPENDENCE ON ω

- Plotting the relative number of points as a function of ω on the same graph reveals a recurrent motif of transitions.

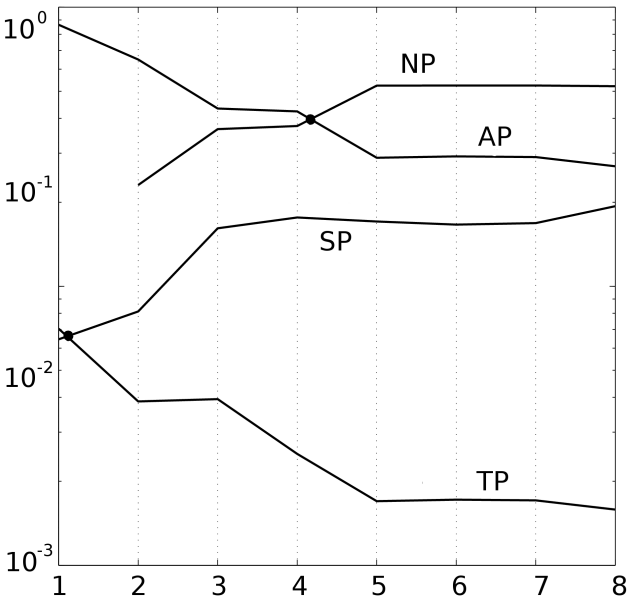
Family 1.1.1.3



Family 1.8.1.1

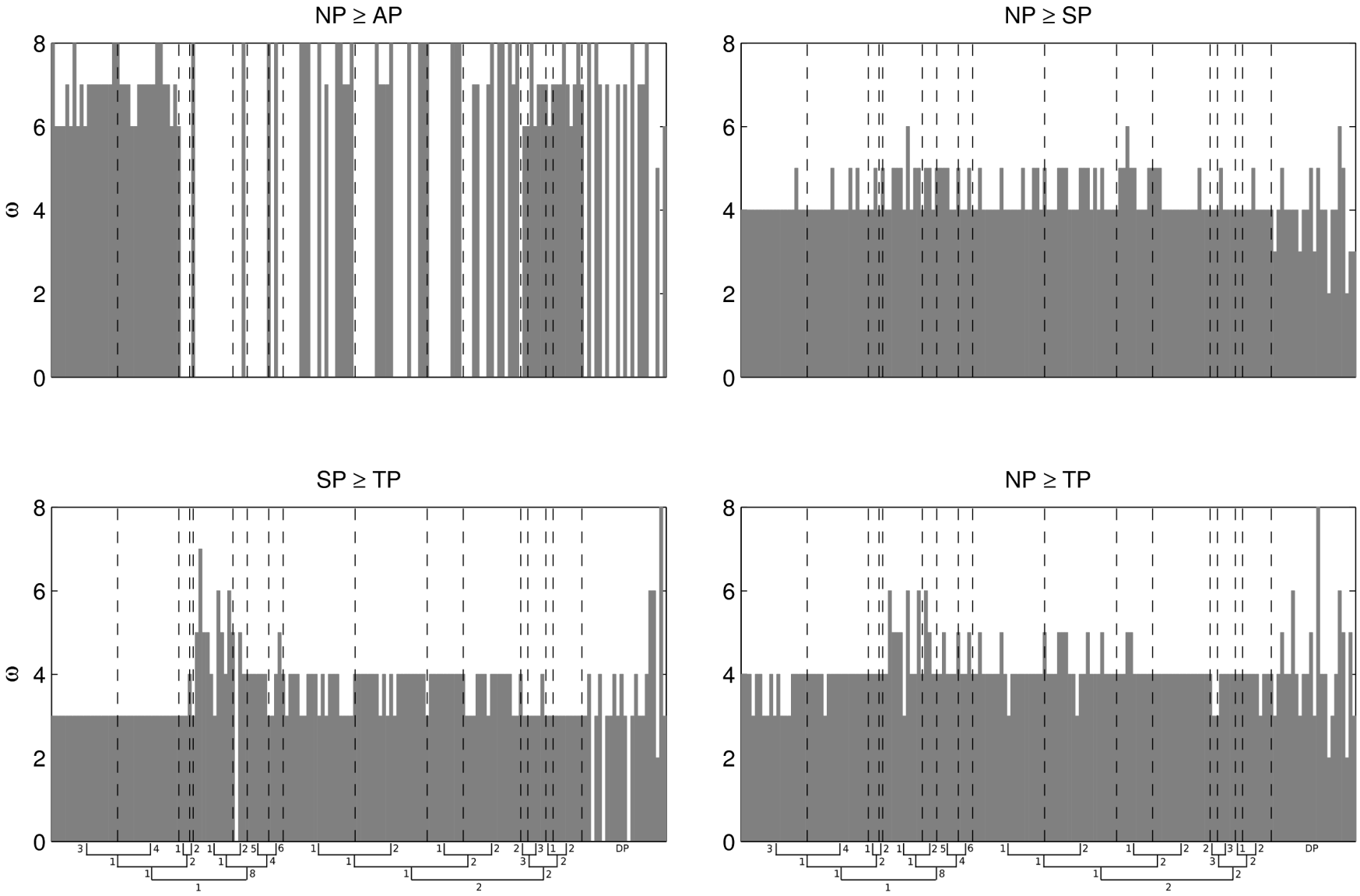


DisProt 34



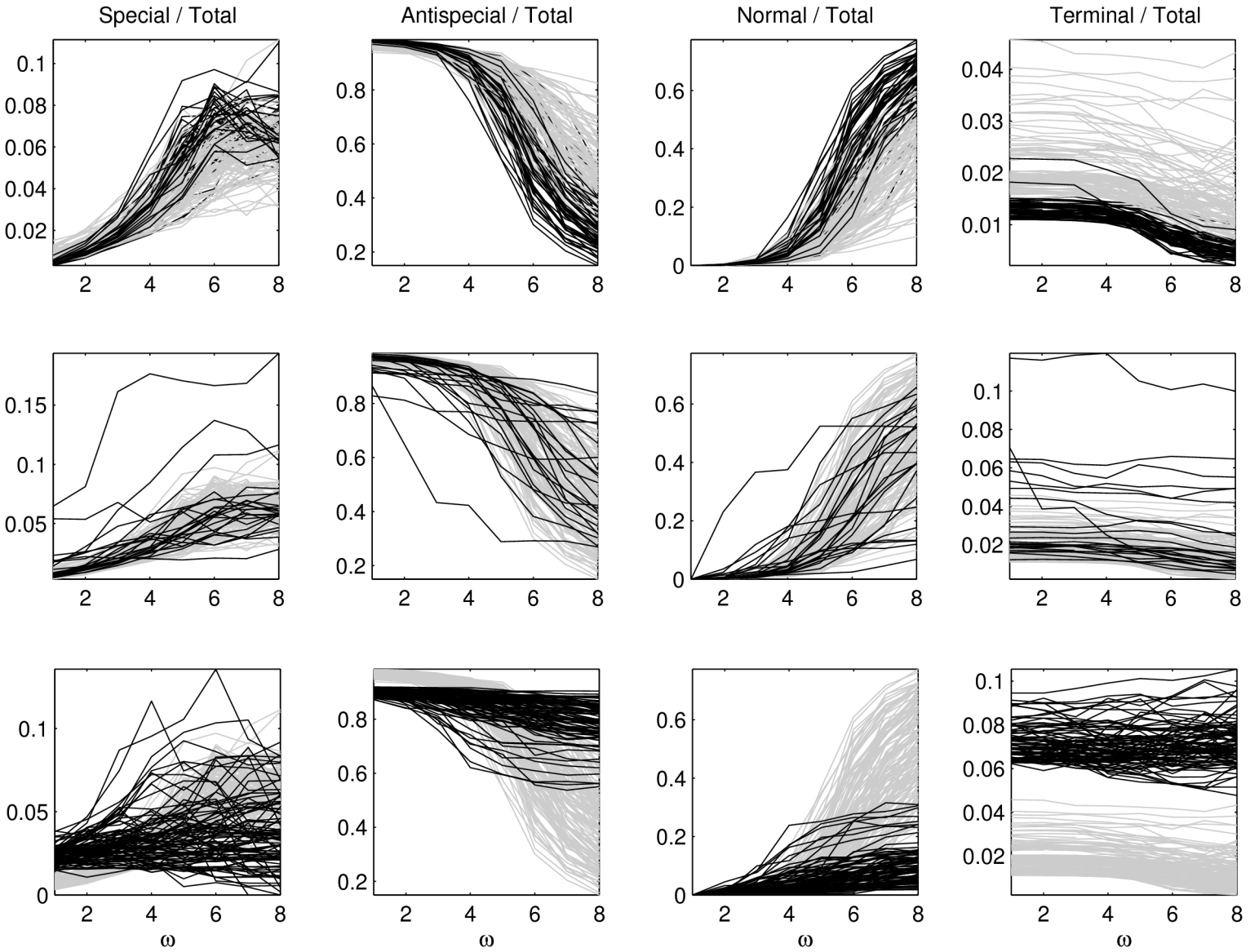
2926606920439054397967548375833719159919910473895631392945311389708800463637987473929840325809307030010774766185
391004135689734801234068539673581567775247599312241880503731683150081948935872832347373134284011181695810360683 6

TRANSITION POINTS IN THE DATASET



2926606920439054397967548375833719159919910473895631392945311389708800463637987473929840325809307030010774766185
3910041356897348012340685396735815677752475993122418805037316831500819489358728323473731342840111816958103606836

SINGLE MEASURES vs ω

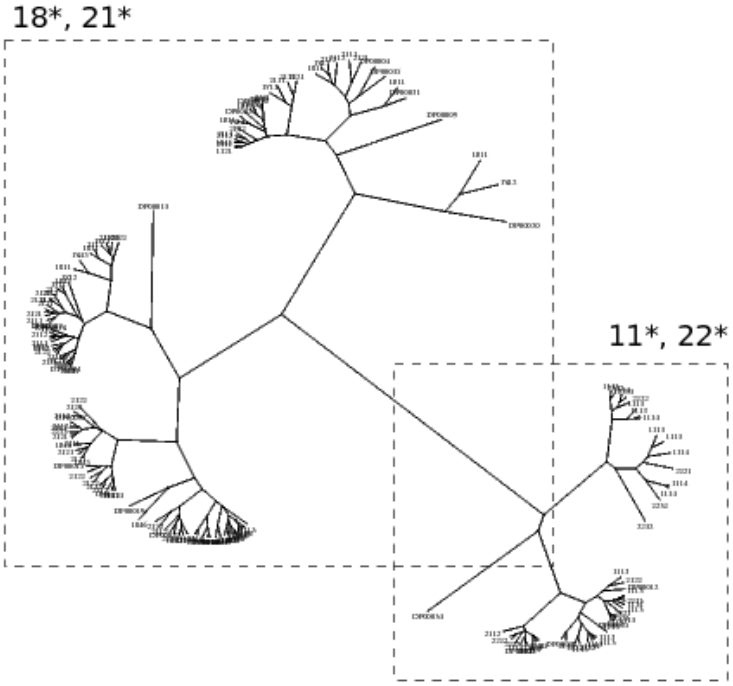


Black: (1.1,2.2)
Grey: (1.8,2.1)

Black: D_2
Grey: D_1

Black: D_3
Grey: D_1

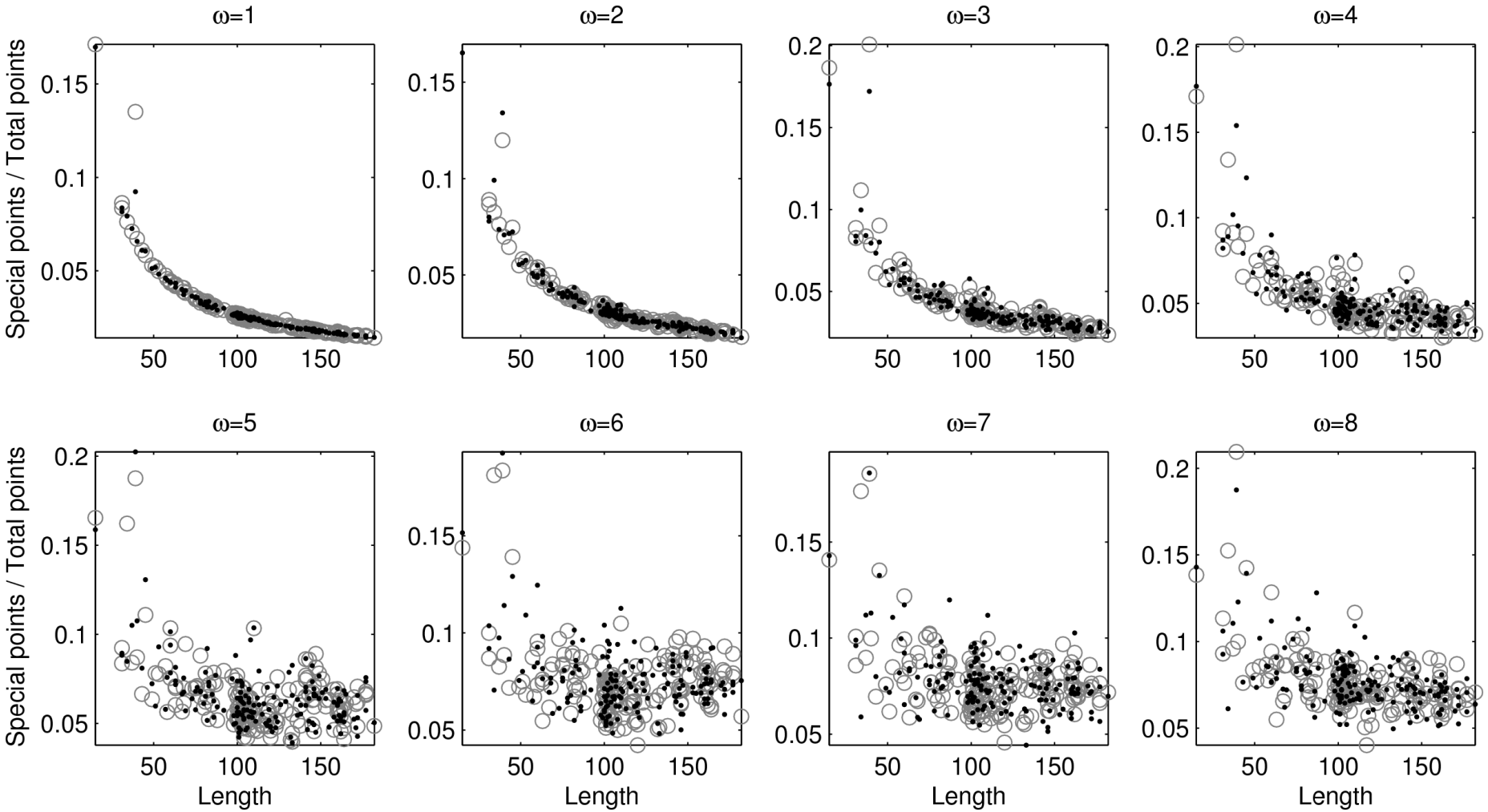
LAWS GOVERNING RANDOM PERMUTATIONS OF POLYPEPTIDES



- Loci are not encoded in the sequence
- Random permutations of most polypeptides amass along linear loci
- The shape of these loci depends on ω
- Proteins rarely escape the loci of their random permutations

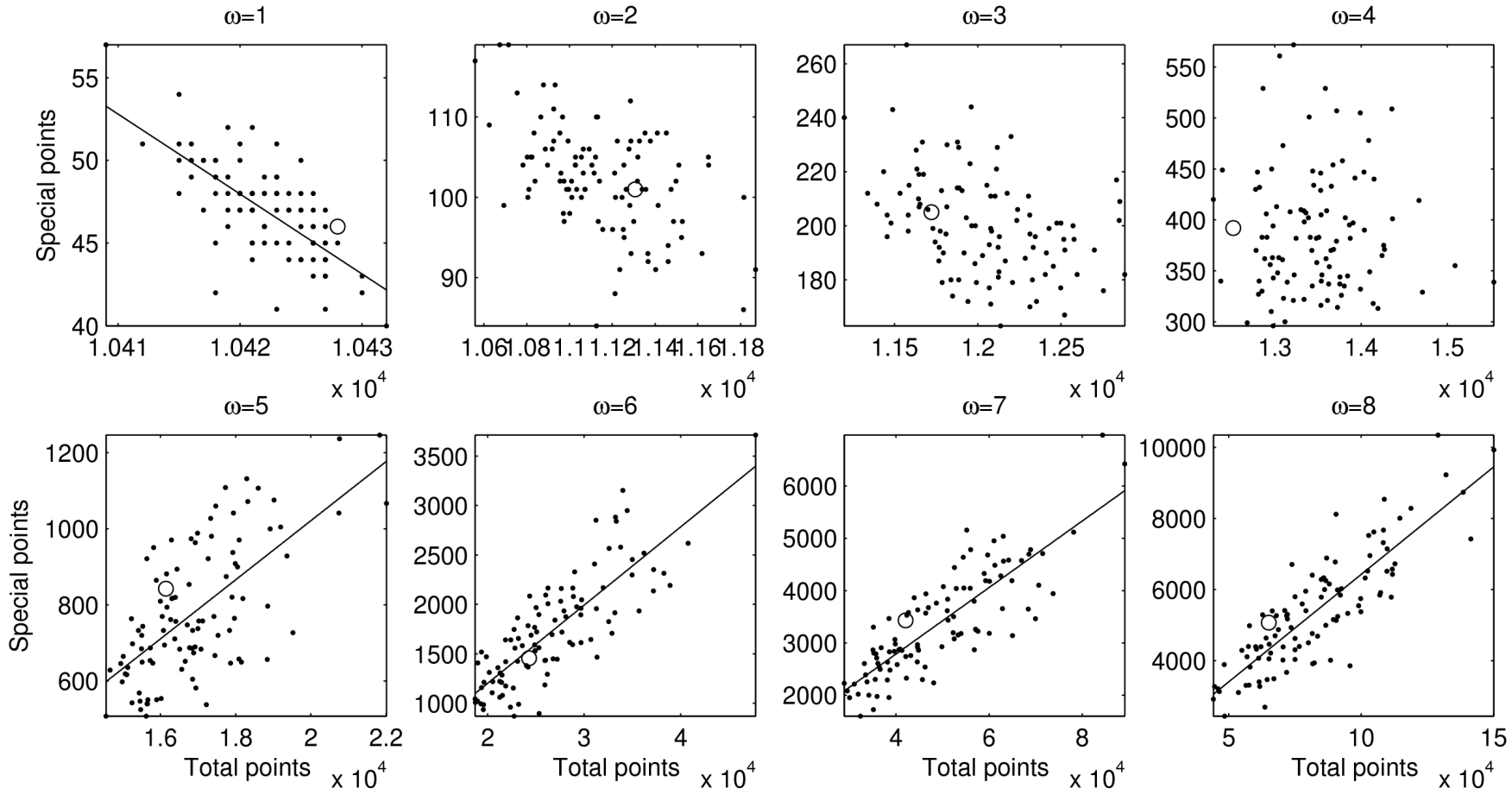
2926606920439054397967548375833719159919910473895631392945311389708800463637987473929840325809307030010774766185
3910041356897348012340685396735815677752475993122418805037316831500819489358728323473731342840111816958103606836

LOCI ARE NOT ENCODED IN THE SEQUENCE

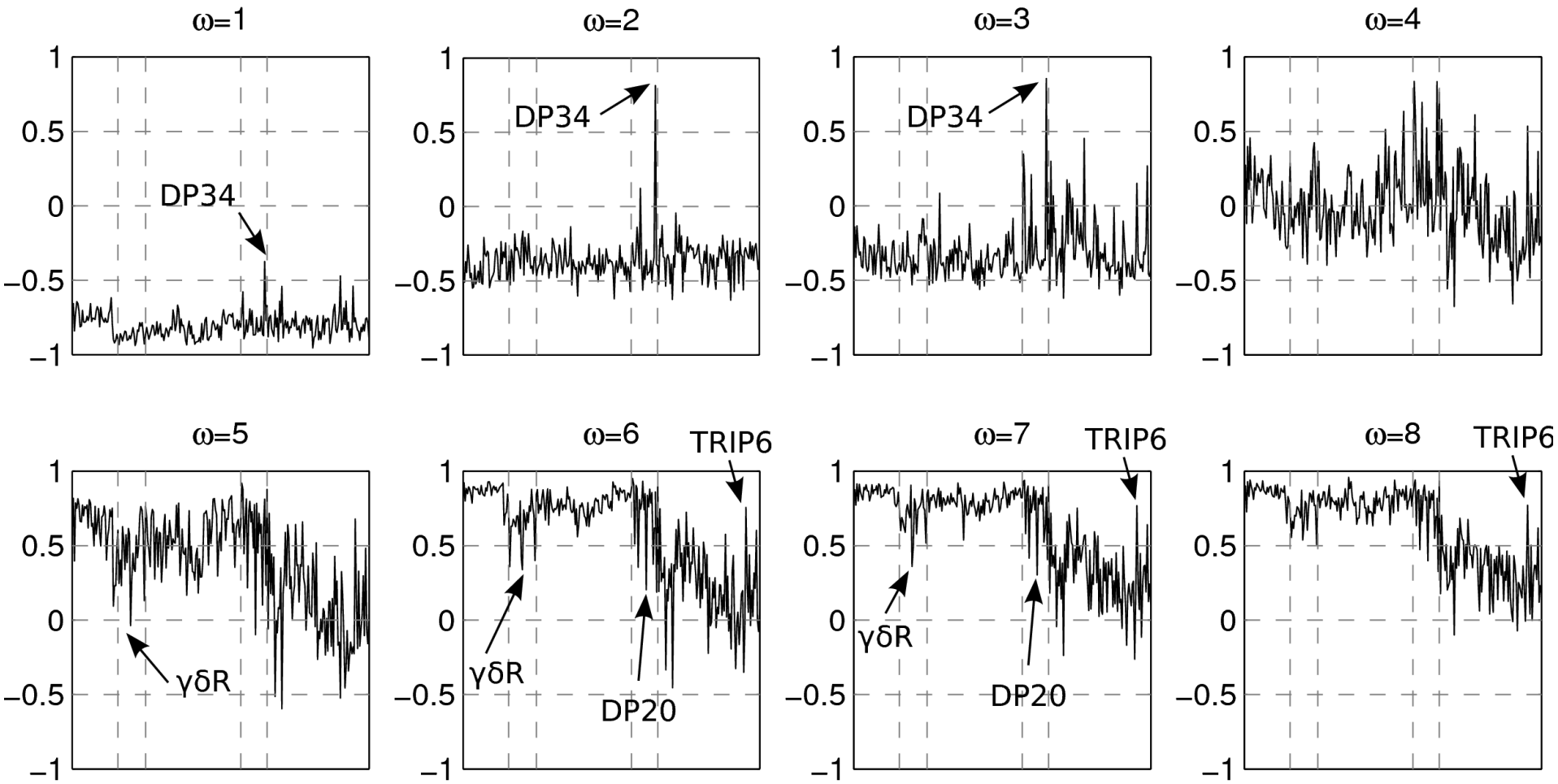


2926606920439054397967548375833719159919910473895631392945311389708800463637987473929840325809307030010774766185
3910041356897348012340685396735815677752475993122418805037316831500819489358728323473731342840111816958103606836

THEN, WHAT IS THE ROLE OF SEQUENCE ON MEASURES?

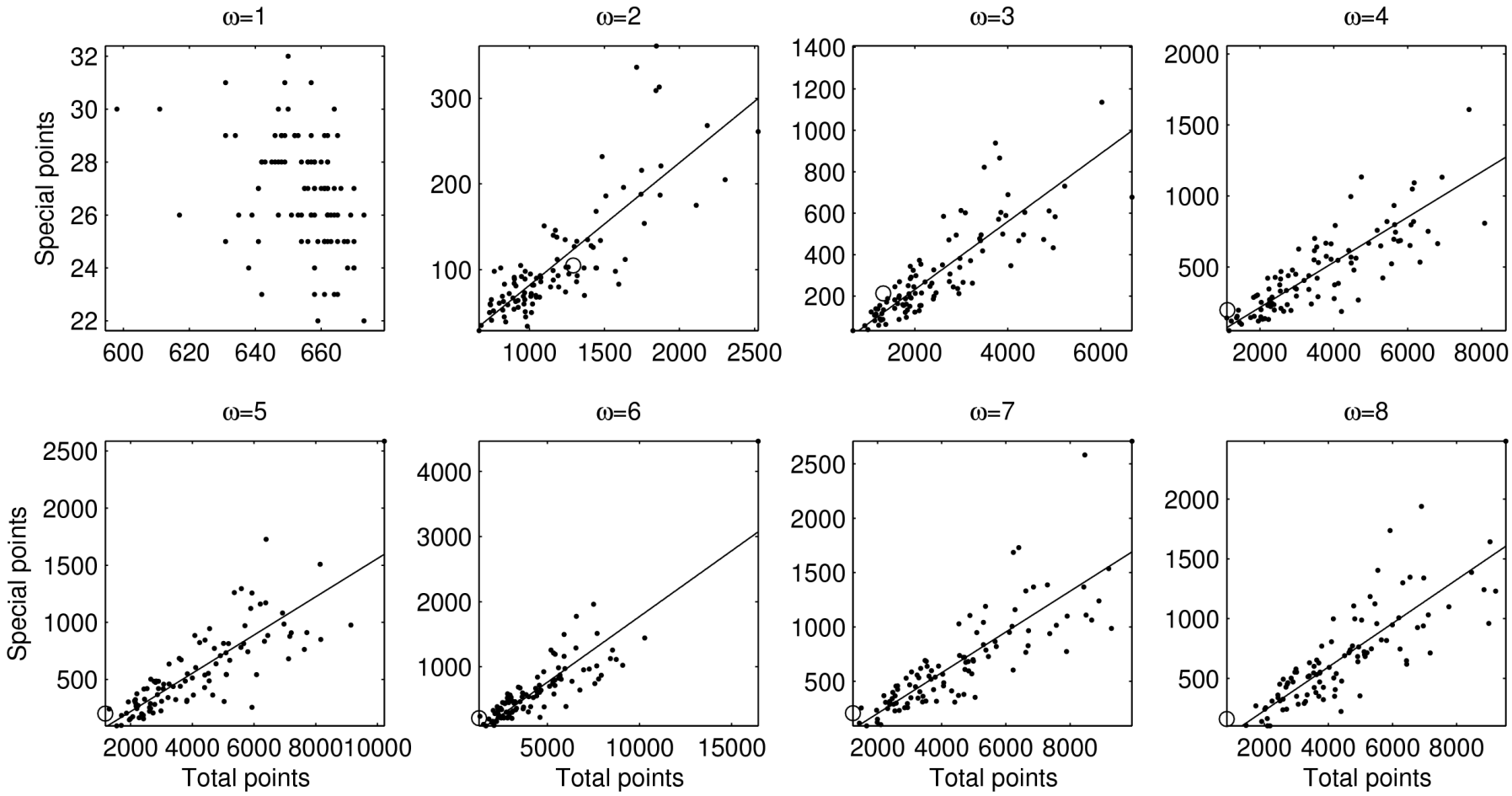


SUPPORT IN THE DATASET



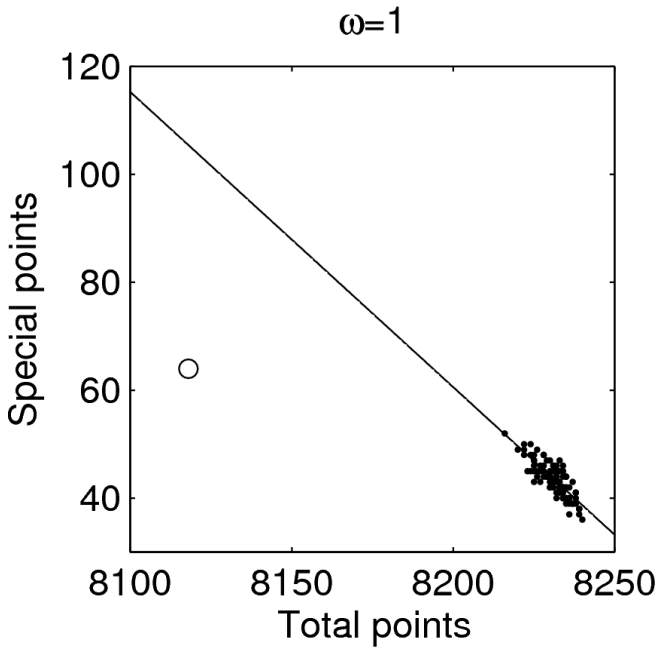
2926606920439054397967548375833719159919910473895631392945311389708800463637987473929840325809307030010774766185
3910041356897348012340685396735815677752475993122418805037316831500819489358728323473731342840111816958103606836

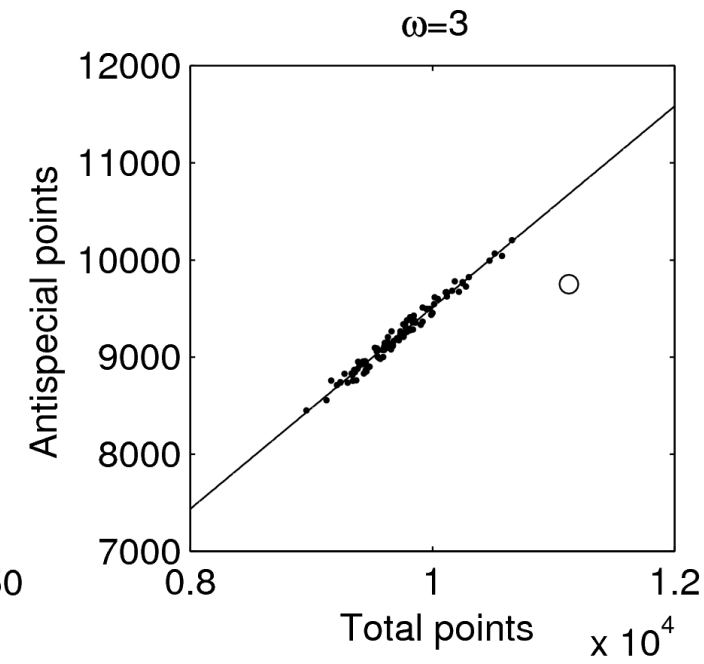
OUTLIERS: DISPROT 34

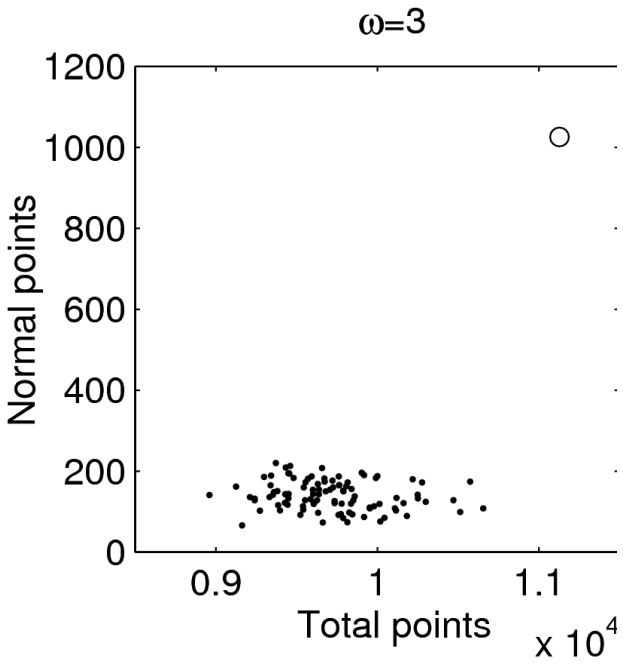


2926606920439054397967548375833719159919910473895631392945311389708800463637987473929840325809307030010774766185
391004135689734801234068539673581567775247599312241880503731683150081948935872832347373134284011181695810360683 6

OUTLIERS: DISPROT 25







2926606920439054397967548375833719159919910473895631392945311389708800463637987473929840325809307030010774766185
3910041356897348012340685396735815677752475993122418805037316831500819489358728323473731342840111816958103606836

CONCLUSIONS

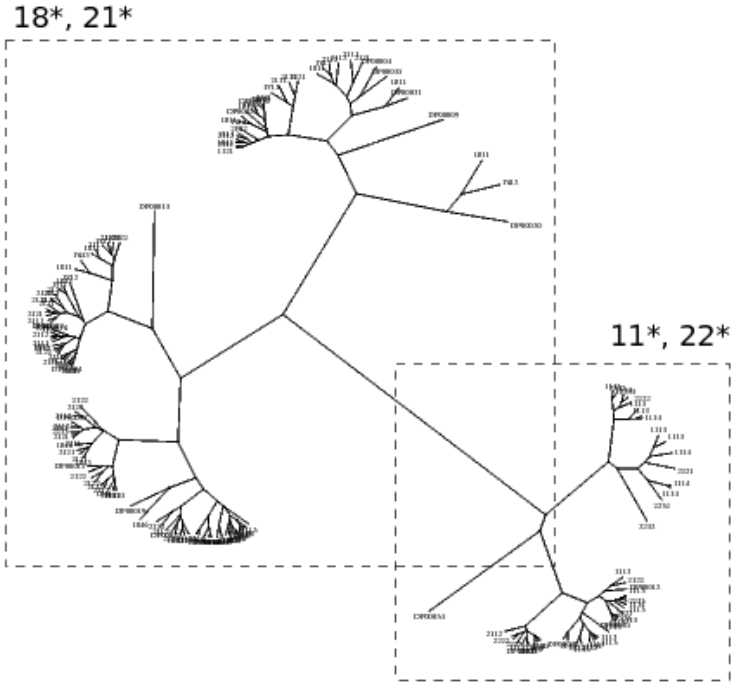
- Natural measures on the abundance of points, arcs and subsequences in suffix graphs grasp structural/functional information of polypeptides.
- Suffix graph measures depend on string length and on ω under a specific set of rules, shared by structurally and functionally diverse polypeptides.
- Rules are influenced by the distribution of symbols more strongly than by the sequence.
- Randomly permuting the sequence of most polypeptides does not allow to escape linear loci whose shapes depend on ω.
- Counterexamples show that none of these rules is unavoidable.

Rules are common to structurally/functionally diverse polypeptides.

- They could capture organizational constraints that cross protein families and are related to stability [Han and Baker, 1995].
- They could capture properties of the ancient peptide world [Qi et al. 2004], or of the rules of assemblage of ancient peptides into domain [Lupas, 2001].
- They could reflect biases and optimizations in the translation machinery [Dufton, 1997].
- Or they could just be the image of constraints in the genome.

THE SUBSEQUENCE COMPOSITION OF POLYPEPTIDES

FABIO CUNIAL | ALBERTO APOSTOLICO



QUESTIONS?

2926606920439054397967548375833719159919910473895631392945311389708800463637987473929840325809307030010774766185
391004135689734801234068539673581567775247599312241880503731683150081948935872832347373134284011181695810360683 6