Proteome Analysis Based on Motif Statistics

Pierre Nicodème*
CNRS - Génopole Evry
523 Place des Terrasses
91000 Evry
France
nicodeme@genopole.cnrs.fr

Tobias Doerks
EMBL-Heidelberg
Meyerhofstrasse 1
69117 Heidelberg
Germany
doerks@EMBL-Heidelberg.de

Martin Vingron
MPI Molecular Genetics
Ihnestrasse 73
14195 Berlin
Germany
vingron@molgen.mpg.de

June 5, 2002

Abstract

Motivation: Even for the amino acid motifs collected in the Prosite database there may be chance occurences as opposed to those occurences where the motif is involved in fold or function of a protein. With recent mathematical advances in assessing the significance of observing such a motif a particular number of times, we can now study the over- or underrepresentation of particular motifs in a complete genome and attempt to make functional deductions.
Results: We demonstrate that statistical over- or under-representation of motifs in complete proteomes may be an indicator of whether, in that organism, we are looking at chance occurrences of the motif or whether the occurrences are sufficiently numerous to suggest a systematic, and thus functionally important occurrence. This has important implications on databank annotations.
Availability: The complete dataset comprising the plotted statistics of 266 Prosite motifs on 42 proteomes is available at `http://algo.inria.fr/nicodeme/proteomes/proteocomp.html`. The software programs used to compute these data have been described in [Nic00, Nic01]. They are available either by Web access as mentionned in these articles, either by direct request to Pierre Nicodème.
Contact: `nicodeme@genopole.cnrs.fr`

*To whom correspondence should be addressed. (Tel: 0033(0)160873800 - Fax: 0033(0)160873809)

1 Introduction

Many protein domains or families can be characterized by short amino acid motifs comprising conserved and variable positions, possibly with spacer regions in between. Such motifs are collected in the Prosite database [HBFB99]. Some Prosite patterns occur very frequently, like, e.g., the so-called P-loop which is given by Alanine or Glycine, then any four residues, then Glycine, followed by Lysine, and lastly Serine or Threonine. In Prosite syntax this is denoted "[AG]-x(4)-G-K-[ST]". More intricate patterns are observed much more rarely than this abundant one. In this article, we study the number of occurrences of Prosite motifs in complete proteomes. Utilizing recently developed mathematical tools to compute how frequently a particular motif is expected to occur, we suggest a biological interpretation of statistical over- or underrepresentation of a motif in the proteome of an organism.

A pattern as frequent as the P-loop is unlikely to be functional wherever it is found in a protein. The cell will recognize the instruction contained in the motif only when the amino acids occur in the right conformation on the surface of the protein. Thus, when one counts occurrences of motifs in proteins one may expect to observe the sum of random occurrences plus the "functional" occurrences. By looking at protein sequences as random text we estimate the level at which we expect to find a particular pattern by chance. As a consequence, our study searches for systematic deviations from the chance expectation.

Recent statistical work on words and on matches of regular expressions, that is reviewed below, provides us with estimates of how often we expect to observe a particular motif by chance alone. Furthermore, it lets us judge whether a motif is significantly under- or overrepresented in a set of proteins. By under/over-representation we mean that the motif occurs significantly more rarely, or more often, than expected by chance. When one specifies the set of proteins under study to be the complete set of gene-products of a particular organism, i.e. its proteome, one can assess for any motif whether *in that organism* it occurs in a quantity that is expected by chance, or whether it is over- or underrepresented. Thus, the focus on a complete proteome lays the basis to assess whether one looks at the consequence of a biological function in that organism or merely at chance occurrences.

Inspection of initial results along those lines led to some remarkable findings. There are indeed motifs which in some organisms are highly over-represented while in other organisms they occur as expected. While in many cases, average occurrence may mean a count of zero, in other cases it will be a positive count. This, in turn, might be taken to be evidence for the associated function or fold to be present in that organism. However, the fact that in other organisms the same motif occurs much more frequently than expected might indicate that it is only functional in those organisms, whereas in an organism where the number of occurrences is low we are simply looking at a chance product. If this holds true, statistical over-representation of a motif in a complete genome would provide us with a clue towards functionality of that motif. It would allow us to discern chance occurrence within one organism based on the over-representation in another one. This aspect seems to be of sufficient interest to merit a dedicated study, on which we want to report here.

2 Motif Statistics

2.1 State of the art

From a mathematical viewpoint, the Prosite motifs describe finite languages (set of words) that are themselves a subclass of the (possibly infinite) regular languages. Therefore, motif statistics belong to the field of word statistics. In a pioneer work, Guibas and Odlyzko [GO81] introduced the autocorrelation polynomial in the case of patterns defined by one word. This polynomial describes the self-overlapping structure of a word. They considered the probability of match of one word in a random text, or the waiting time for the first match. They used univariate generating functions to modelize these problems. The next step was to study the number of occurrences of words (or set of words) in random texts. Pevzner and *al.* [PBM89] consider patterns allowing fixed length gaps of don't care symbols. During the last recent years, Régnier [Rég98, Rég00] and Régnier and Szpankowski [RS98a] study finite sets of words in Bernoulli or Markov texts, proving (1) Gaussian asymptotic laws and (2) Poisson distributions when the number of occurrences is $O(1)$. They give closed formulae for the expectation and the variance of the number of matches in the Bernoulli and Markov cases. The mathematical tool used there is the asymptotic analysis of bivariate generating functions, where one variable counts the length of the random text, and the second variable counts the number of occurrences of the motif. Sinha and Tompa [ST00] give a filtering method for exhaustive search of exceptional patterns of fixed length including multiple choice letters. Schbath and *al.* [SPdT95], Prum and *al.* [PRdT95], and Reinert and Schbath [RS98b] study by probabilistic methods words with unexpected frequencies and multiple words. Using Poisson approximation and Chen-Stein method, Reinert and Schbath [RS98b] showed that the number of occurrences of non overlapping finite sets of words is asymptotically Poisson, the total deviation error being $O(1/n)$ for texts of size n. Nicodème and *al.* [NSF01] consider the case of matches with any regular expression. Using marked automata that recognize random texts where a mark is added after each match, and applying the Chomski-Schützenberger algorithm, they compute the bivariate generating function counting the number of matches. From there, they give algorithms to compute exactly or asymptotically the expectation and the variance of the number of matches, and they prove an asymptotic normal distribution. Nicodème [Nic01] considers the subclass of non self-overlapping motifs in the Bernoulli model; in this latter case, he proves a Poisson law for rare occurrences in large texts and gives a closed formula for the expectation. He verifies numerically that these results apply with a good approximation to motifs whose expectation is not too large. This is induced by the fact that, when the probability of occurrence of a motif is small, the probability of self-overlapping is much smaller.

2.2 Sketch of the statistical method

We describe briefly here our statistical and algorithmic approach. For full details, we refer to Nicodème [Nic01] and to Nicodème *et al.* [NSF01] for the approximate and exact computation methods, respectively.

The probabilistic model for the texts is the non-uniform Bernoulli model.

We consider in this section the bivariate generating function

$$F(z,u) = \sum_{n\geq 0, k\geq 0} f_{n,k} u^k z^n = \sum \mathbf{P}(X_n = k) u^k z^n, \tag{1}$$

where X_n is the random variable counting the number of matches with the considered motif in a (random) text of size n in the renewal context. Let $\mu_n = \mathbf{E}(X_n)$ denote the corresponding expectation and $[z^n]\phi(z)$ denote as usual the n^{th} Taylor coefficient of a function $\phi(z)$. We have

$$\mu(z) = \sum_{n\geq 0} \mu_n z^n = \left.\frac{\partial F(z,u)}{\partial u}\right|_{u=1} = \frac{P(z)}{(1-z)^2 Q(z)}, \tag{2}$$

$$\text{and} \qquad \mu_n = [z^n]\frac{P(z)}{(1-z)^2 Q(z)} \quad \asymp \quad \alpha \times n + \beta \quad \text{for large } n, \tag{3}$$

where $P(z)$ and $Q(z)$ are polynomials such that $P(1)Q(1) \neq 0$, and α and β are functions of $P(1), Q(1)$, $P'(1)$ and $Q'(1)$. Equation 2 is a consequence of the presence of the dominant pole $z = 1$ in $F(z,1)$. It follows from Perron-Frobenius theory for non-negative matrices that all the moduli of the roots of the equation $Q(z) = 0$ are larger than 1. The asymptotic value of μ_n is obtained by singularity analysis in the neighborhood of $z = 1$.

Approximate computations. For non self-overlapping motifs, there is a closed formula for the bivariate generating function counting the number of matches. Consider a motif M. Let

$$M(z) = \sum_{w\in M} \pi_w z^{|w|},$$

where π_w and $|w|$ respectively are the probability and the number of letters of the word w. Let $N = \Sigma^\star - \Sigma^\star M \Sigma^\star$ be the language of texts over an alphabet Σ that are not matched by the motif M. Parsing the texts by the occurrences of the motif, we have the equations

$$\Sigma^\star = \bigcup_{k\geq 0} N(MN)^k \quad \text{and} \quad \Sigma^\star_m = \bigcup_{k\geq 0} N(MmN)^k,$$

where $\Sigma^\star_m$ is the union of all the texts of $\Sigma^\star$ where a mark m has been inserted after each match with M. This translates to the following two equations on generating functions,

$$\frac{1}{1-z} = \frac{N(z)}{1-N(z)M(z)} \quad \text{and} \quad F(z,u) = \frac{N(z)}{1-uN(z)M(z)}.$$

Eliminating $N(z)$ between the two equations gives

$$F(z,u) = \frac{1}{1-z+(1-u)M(z)}$$

and from there we obtain the expectation. Taking the Taylor expansion of $F(z,u)$ in the neighborhood of $u = 0$, this gives after some analysis that, asymptotically, rare occurrences obey a Poisson law,

$$f_{n,k} = \frac{\rho^{-n-1}}{1-M'(\rho)} \frac{(\alpha n)^k}{k!} \left(1 + O\left(\frac{1}{n}\right)\right), \qquad \alpha = \left(\frac{\rho^{-1}M(\rho)}{1-M'(\rho)}\right).$$

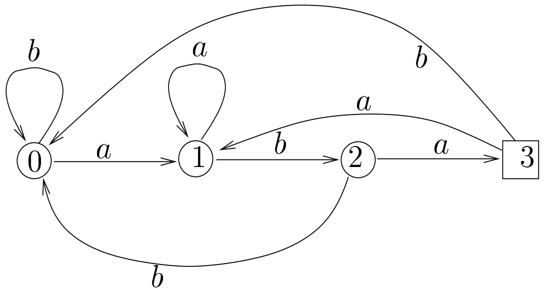


Figure 1: Automaton recognizing all matches with aba (renewal context).

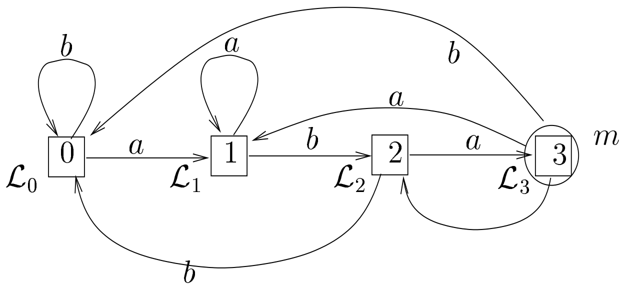


Figure 2: Marked automaton recognizing marked texts where a mark m is added after each match aba in every text (renewal context).

In this last equation, ρ is the only root of the equation $1 - z + M(z) = 0$ inside the disk centered at the origin and of radius 2. This root is real and larger than 1. As mentioned in Section 2.1, for most motifs, the self-overlapping structure is weak, and the results (expectation, Poisson law) computed without taking in account the overlap structure are good approximations. However, this does not work for motifs whose occurrence probability is large (inducing often a large probability of self-overlap). We note that these motifs have large expectations and use for them the exact computation described in the following section.

Exact computation. The exact computation uses an automaton construction. We refer to standard textbooks such as Kelley [Kel95] or Kozen [Koz97] for the definitions regarding automata. As an example, we present the computation for the motif aba on the two letter alphabet $\Sigma = \{a, b\}$. The method generalizes to any regular expression over a finite alphabet. The (renewal) DFA A of Figure 1 is deduced from a DFA recognizing the language $\Sigma^\star aba$ by moving the transitions from the terminal states to the states accessed with the same letter from the start state. The (renewal) marked DFA A_m of Figure 2 recognizes texts where a mark m is inserted after each match with aba. In these automata, the initial state is state 0, and the final states are represented by a square. The large circle of state 3 of A_m means that this state is marked. Entering this state with any transition implies the insertion of the letter m. The Chomski-Schützemberger algorithm works as follows. We consider the deterministic automaton A_m from Figure 2. For $j \in \{0, 1, 2, 3\}$, we introduce the language $\mathcal{L}_j$ of all words recognized by the automaton with start state (j) and final state (3). The languages are connected by the system of formal equations

$$\mathcal{L}_0 = a\mathcal{L}_1 + b\mathcal{L}_0 + \epsilon, \quad \ldots \quad , \mathcal{L}_2 = b\mathcal{L}_0 + am\mathcal{L}_3 + \epsilon, \tag{4}$$

where ϵ is the empty word. These equations translate to the linear system on generating functions

$$L_0 = aL_1 + bL_0 + 1, \quad \ldots \quad , L_2 = bL_0 + amL_3 + 1, \tag{5}$$

whose solution gives (with probability π_a and π_b for letters a and b)

$$L_0(a, b, m) = \frac{1 + ab}{1 - a - b + ab - ab^2 - a^2bm} \quad \text{and} \quad F(z, u) = L_0(\pi_a z, \pi_a z, u) = \sum_{n,k} f_{n,k} u^k z^n, \tag{6}$$

where $[a^i b^j m^k]L_0(a,b,m)$ counts the number of texts with i letters a, j letters b and k matches and $f_{n,k}$ is the probability that a text of size n contains k matches. From there, we obtain the expectation for texts of size n as mentionned in Equation 2.

3 Calibration to correct for different genome size

In the context of the current work, the theory summarized above shall be employed to compare the abundance of a certain motif across different organism. To obtain an intuitive measure of significance we first map the $p-$values into a standard normal such that one can meaningfully describe abundance in terms of standard deviations below/above the mean. Let us consider $X_{\mathcal{P}_\mu}$, a Poisson random variable of expectation μ, and $Y_\mathcal{N}$ a random variable with standard Gaussian distribution. Let ω_n represent the number of observations of a motif in a proteome of size n, and $p = \mathbf{P}(X_{\mathcal{P}_\mu} \geq \omega_n)$ if $\omega_n > \mu$ and $p = \mathbf{P}(X_{\mathcal{P}_\mu} \leq \omega_n)$ if $\omega_n < \mu$. We define the statistical calibration γ_n of these observations by $\gamma_n = [\,\text{sign of}(\omega_n - \mu)] \times x$ where x is the root of the equation $\mathbf{P}(Y_\mathcal{N} > x) = p$. We thus scale a p–value computed on a Poisson distribution onto a Gaussian distribution. We do not use here a Poisson approximation method (such as used by Chen and Stein), but apply the results proving that the Poisson distribution is a good approximation of the distribution of number of occurrences for most motifs and proteomes of hundreds of thousands of amino-acids.

A further problem arises from comparing calibrations for proteomes with very different length. The proteome of *Arabidopsis* is about seventy times larger than the proteome of *Mycoplasma genitalium*. The calibration described above is not stable when expectations and observations are scaled up or down linearly. A Z–score $Z = \frac{\omega-\mu}{\sqrt{\mu}}$ would not be stable either. A helpful and plausible assumption would say that the number of occurrences of most Prosite motifs is linear with the size of the proteome. We accept this as a basis for our genome comparisons and thus scale every proteome to a "standard proteome size" of 1 million amino-acids. Thus, we define $\mu_s = \frac{1000000}{n}\mu_n$, $\omega_s = \frac{1000000}{n}\omega_n$, where n is the size of the proteome. From μ_s and ω_s we compute γ_s. See Table 4 for examples.

PS00010	C-x-[DN]-x(4)-[FY]-x-C-x-C
PS00016	R-G-D
PS00027	[LIVMFYG]-[ASLVR]-x(2)-[LIVSTACN]-x-[LIVM]-x(4)-[LIV]-[RKNQESTAIY]-[LIVFSTNKH]-W-[FYVC]-x-[NDQTAH]-x(5)-[RKNAIMW]
PS00028	C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H
PS00039	[LIVMF](2)-D-E-A-D-[RKEN]-x-[LIVMFYGSTN]
PS00043	[LIVAPKR]-[PILV]-x-[EQTIVMR]-x(2)-[LIVM]-x(3)-[LIVMFYK]-x-[LIVFT]-[DNGSTK]-[RGTLV]-x-[STAIVP]-[LIVA]-x(2)-[STAGV]-[LIVMFYH]-x(2)-[LMA]
PS00044	[NQKRHSTAG]-[LIVMFYTA]-x(2)-[STAGLV]-[STAG]-x(4)-[LIVMYCTQR]-[PSTANLVER]-x-[PSTAGQV]-[PSTAGNVMF]-[LIVMFA]-[STAGH]-x(2)-[LIVMF]-x(2)-[LIVMFW]-[RKEAV]-x(2)-[LIVMFYNTAE]-x(3)-[LIMVT]

Table 1: Prosite description of the motifs used in Figure 3.

	Taxo.	Abbreviation	Proteome description	Length
Eukaryotes		dros-mela	Drosophila melanogaster	6617803
		caen-eleg	Caenorhabditis elegans	7863058
		sacc-cere	Saccharomyces cerevisiae	2937205
		arab-thal	Arabidopsis thaliana	11304307
Archae-Bacteria		aero-pern	Aeropyrum pernix K1	638811
		arch-fulg	Archaeoglobus fulgidus	660832
		meth-jann	Methanococcus jannaschii	502362
		meth-ther	Methanobacterium thermoautotrophicum	525164
		pyro-abys	Pyrococcus abyssi	535786
		pyro-hori	Pyrococcus horikoshi	568584
		halo-nrc1	Halobacterium sp. NRC-1	624005
		ther-acid	Thermoplasma acidophilum	453115
Bacteria	Ae	aqui-aeol	Aquifex aeolicus	488484
	Fb	baci-halo	Bacillus halodurans	1169204
	Fb	baci-subt	Bacillus subtilis	1218487
	Fb	myco-geni	Mycoplasma genitalium	175729
	Fb	myco-pneu	Mycoplasma pneumoniae	237930
	Fb	urea-urea	Ureaplasma parvum	227717
	Fb	lact-lact	Lactococcus lactis (subsp. lactis) strain IL1403	655989
	Fa	myco-lepr	Mycobacterium leprae	520057
	Fa	myco-tube	Mycobacterium tuberculosis strain H37Rv	1317198
	STT	borr-burg	Borrelia burgdorferi	352644
	STT	trep-pall	Treponema pallidum	349913
	STT	ther-mari	Thermotoga maritima	582037
	STT	dein-radi	Deinococcus radiodurans	951455
	PCV	chla-muri	Chlamydia muridarum	324428
	PCV	chla-pneu	Chlamydia pneumoniae strain CWL029	361707
	PCV	chla-trac	Chlamydia trachomatis	312118
	Pab	rick-prow	Rickettsia prowazekii	279044
	Pab	neis-mena	Neisseria meningitidis strain Z2491 (serogroup A)	582084
	Pab	neis-menb	Neisseria meningitidis strain MC58 (serogroup B)	573863
	Pg	esch-coli	Escherichia coli K-12	1373785
	Pg	buch-aphi	Buchnera aphidicola (subsp. Acyrthosiphon pisum)	187023
	Pg	haem-infl	Haemophilus influenzae	526801
	Pg	past-mult	Pasteurella multocida	667675
	Pg	pseu-aeru	Pseudomonas aeruginosa	1856757
	Pg	xyla-fast	Xylella fastidiosa	744437
	Pg	vibr-chol	Vibrio cholerae	1156096
	Pd	camp-jeju	Campylobacter jejuni	503501
	Pd	heli-2pyl	Helicobacter pylori strain 26695	491768
	Pd	heli-Jpyl	Helicobacter pylori strain J99	493049
	C	syne-cyan	Synechocystis sp. PCC 6803	1027015

Abbreviations used in the taxonomy of the bacteria:

○ Ae: *Aquificales.*

○ *Firmicutes.* ■ Fb: *Firm. Bacillus/Clostridium group.* ■ Fa: *Firm. Actinobacteria.*

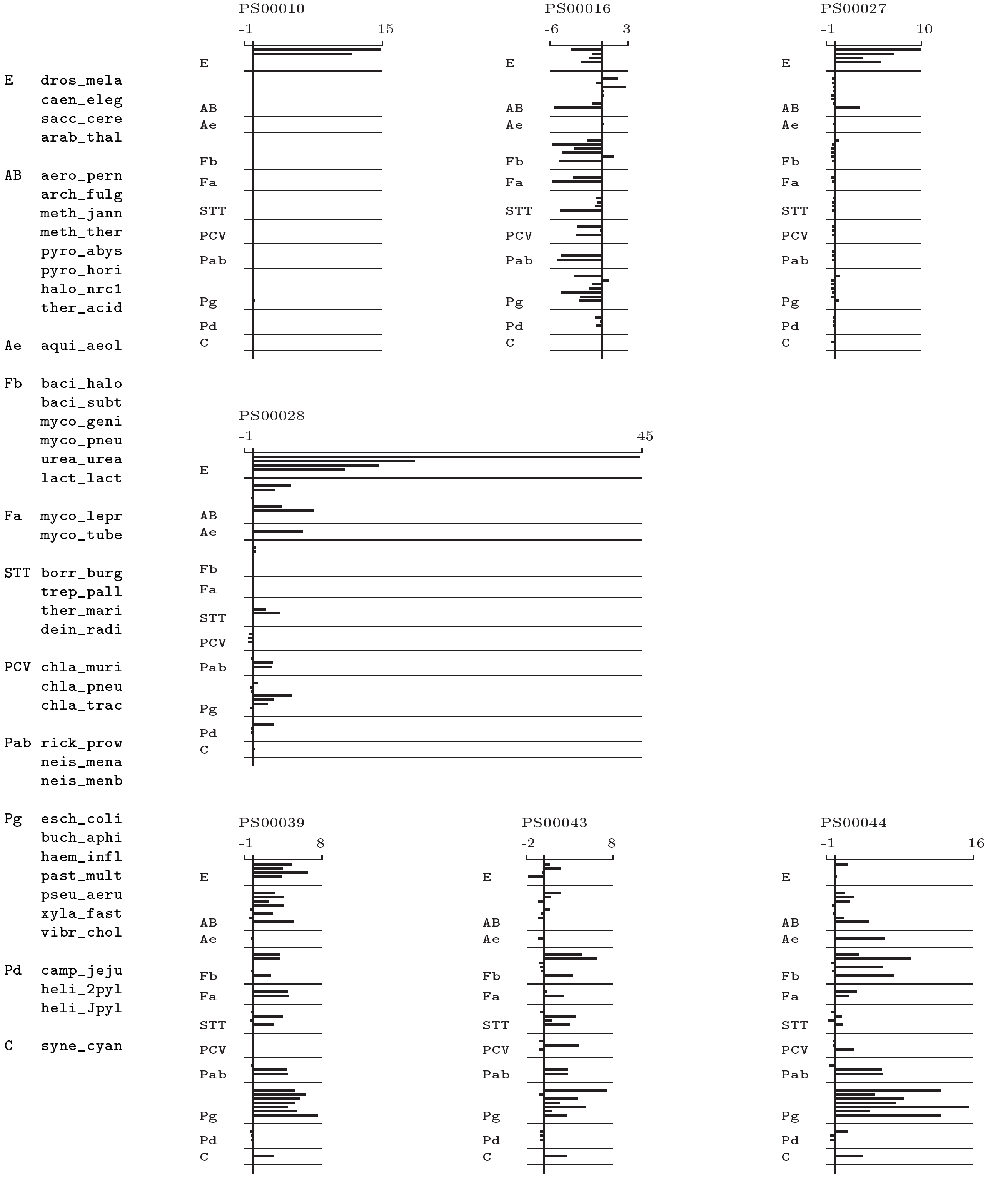
○ STT: *Spirochaetales, Thermotogales, Thermus/Deinococcus group.*

○ PCV: *Planctomyces/Chlamydia/Verrucomicrobium group.*

○ *Proteobacteria.* ■ Pab: *Proteo. alpha and beta.* ■ Pg: *Proteo. gamma.* ■ Pd: *Proteo delta.*

○ C: *Cyanobacteria.*

Table 2:



PS00010 Aspartic acid and asparagine hydroxylation site
PS00016 Cell attachment sequence
PS00027 'Homeobox' domain signature
PS00028 Zinc finger, C2H2 type
PS00039 AD-box subfamily ATP-dependent helicases
PS00043 Bacterial regulatory proteins, gntR family
PS00044 Bacterial regulatory proteins, lysR family

Figure 3: Over- and under-represented characteristic motifs. The scaled statistical calibrations γ_s used in the plots are computed from the scaled expectations μ_s and observations ω_s (See Section 3 and Table 4). See also Table 3 for the unscaled expectations and observations of the motifs.

4 Proteome comparison

Analysis of abundance of Prosite motifs has been carried out for the complete predicted proteomes of 42 organisms listed in Table 1. We focussed on Prosite motifs that were somewhat frequent because otherwise the discrete nature of the observed counts would make it impossible to assess statistical significance in any meaningful way. This results in 266 Prosite motifs that we used for this study. The number of occurrences of each motif in the proteome of each organism was determined and, for each motif, its expected distribution was computed. This allowed to assign to each motif and each proteome the number of standard deviations an observed count of motif occurrences was away from its expectation.

Table 3 in the Appendix gives figures for observed and expected motif occurrences. Table 4 in the Appendix demonstrates the effect of scaling to a common genome size. The plots in Figure 3 are the key visualization tool for the differences in abundance of this subset of motifs in the 42 organisms. Organisms are grouped into Eukaryotes (E), Archeae (AB), and Eubacteria (Ae to C) divided along a taxonomy from NCBI (Table **??** and Figure 4). Each row of a plot corresponds to one organism. Horizontal bars in the plots indicate the number of standard deviations the calibrated observed counts are away from calibrated expectation.

Effect of calibration. The effect of the calibration procedure can seen on the example of the DEAD helicases (Prosite motif PS00039). Helicases, and in particular the subfamily characterized by the sequence motif DEAD/DEAH, are essential in RNA metabolism and processing and thus can be found in all organisms. Inspection of the absolute abundance (see Appendix) of the Prosite motif firstly shows a much larger number of occurrences in the four eukaryote species (20 or more) than in the others, where the motif typically is found just a couple of times per proteome. We interpret this as a consequence of the larger genome size rather than a biological effect that we would want to interpret. The calibration takes into account the genome size and leads to the significances shown in Figure 3. There all bars indicate a certain abundance of the motif, without displaying any particular preference for one or the other kingdom or organism. Note, however, that there are some organisms with a count of zero. This is due to the motif definition being too restrictive to identify the domains in those organisms. A search, e.g., in the SMART database [SCD$^+$00] does identify the helicase domain there as well.

Differences in abundance between species. There are many transcription factor domains known to be specific for Eukaryotes, like the zinc-finger domain, the homeodomain, or the basic leucine zipper domain. Others, like the helix-turn-helix motif are specific for Eubacteria. Figure 3 contains the bar plots for the significance of the abundance of several of these domains. PS00027, the homeobox domain motif, shows clear overrepresentation in the Eukaryotes. In the other kingdoms it is either absent altogether, or, present but statistically revealed as being perfectly within the range of random fluctuations. Likewise, the zinc-finger domain motif (PS00028) is strongly over-represented in Eukaryotes, while the largest significance in any of the other organisms is well below these values. Several bacteria contain the motif, yet at a level that does not seem significant. Basic leucine zipper proteins display the same behavior (data not shown).

The helix-turn-helix motif may serve as an example for a Eubacteria specific motif.

Figure 3 contains plots for two of these motifs, namely the one for the repressor of the gluconate operon (gntR, PS00043) and the lysR family of bacterial activators (PS00044). Both motifs are over-represented by several standard deviations in many of the bacteria whereas the occurrences in Eucaryotes, after scaling to their genome size, are not significant. This is not contradicted by the absence of the motif from some bacteria, which is confirmed by the SMART database [SCD+00].

Based on these observation, one can search the data for further cases of group-specific over-representation. As an example, Figure 3 contains the plot for the PS00010, the pattern for a particular post-translational modification, namely a hydroxylation of Asp or Asn. This patterns occurs only in Eukaryotes where in the worm and in the fly it is strongly over-represented. It is absent in yeast while in Arabidopsis it occurs only at a level that might well be a product of chance. This would suggest, in particular, that the occurrences in the plant might not be functional. Prosite documentation gives the taxonomic range of this motif as "eukaryotic", albeit not explicitly naming any plant occurrence.

Are there "avoided patterns"? In some plots a characteristic under-representation of a particular motif can be observed. The Arg-Gly-Asp cell attachment motif (PS00016 in Figure 3) may serve as an example for this type of behavior. Based on these data, there generally seems to be a strong trend to avoid this motif in a protein sequence. The motif occurs in fibronectin and is crucial for the interaction with its cell surface receptor [RP86]. The motif appears to play a role in cell adhesion in many proteins where it occurs. Interestingly, snake disintegrins which inhibit the binding of ligands to the integrin receptor exhibit the same motif. These data may lead to the speculation, that the motif is actively avoided because chance occurrence might have unwanted effects.

5 Discussion

We provide in this article some novel ideas on proteome analysis and on the significance of occurrences of Prosite motifs in proteomes. The key point lies in the connection between the statistical abundance of a motif in an organism's proteome and the functionality of the corresponding domain. In random text we would observe motifs at a frequency expected by chance and for a non-functional motif, the proteome is like random text into which the motif is embedded. Accordingly, we interpret statistical over-representation - and maybe even under-representation - as in indicator that the motif is in some sense functional.

Several examples were given of motifs that, biologically, are specific to particular organisms and whose abundance was statistically much higher where the motif was functional than where it was not. The mechanism, by which nature achieves this over-representation clearly seems to be evolution, i.e. the re-use of domains and the formation of gene families. On the other hand there appear to be patterns the use of which in protein-text is avoided. The example of the RGD cell adhesion motif was given. Snake venoms contain disintegrins that contain this very motif and which are involved in the inhibition of the interaction of cell adhesion molecules with their receptors. This example suggests that free utilization of the motif at a rate at which such a short pattern would be expected might be disadvantageous and possibly selected against during evolution.

In this study, we have focussed on motifs that are sufficiently frequent to actually be observed. Of course, when a motif does not occur in an organism it is trivially not functional. However, we have presented cases where a motif may exist, and only its low level of abundance in comparison with other organisms hints at the random nature of this occurrence. We do not claim that this alone constitutes proof. It may certainly only be judged as a hint that requires experimental testing.

Sometimes we do not identify a domain because the Prosite pattern is not fully conserved and it would require, e.g., a hidden Markov model to identify the domain. A natural continuation of this work would thus be the study of the abundance of occurrences of HMM hits in fully sequences organisms.

The complete dataset comprising the statistics of 266 Prosite motifs on 42 proteomes is available at `http://algo.inria.fr/nicodeme/proteomes/proteocomp.html`. The HTML file provides hyperlinks to Prosite and pointers to Postscript plots. For each motif, a plot comparing the proteomes and for each pair of proteomes, a plot comparing the calibrations of the 266 motifs is available. These data will be regularly updated as new proteomes become available.

References

[GO81] L. J. Guibas and A. M. Odlyzko. String overlaps, pattern matching, and nontransitive games. *Journal of Combinatorial Theory. Series A*, 30(2):183–208, 1981.

[HBFB99] K. Hofmann, P. Bucher, L. Falquet, and A. Bairoch. The PROSITE database, its status in 1999. *Nucleic Acids Res.*, 27:215–219, 1999.

[Kel95] Dean Kelley. *Automata and formal languages.* Prentice Hall Inc., Englewood Cliffs, NJ, 1995. An introduction.

[Koz97] Dexter C. Kozen. *Automata and computability.* Springer-Verlag, New York, 1997.

[Nic00] Pierre Nicodème. Regexpcount, a symbolic package for counting problems on regular expressions and words. In *German Conference on Bioinformatics GCB2000*, pages 63–73, 2000. Heidelberg, Oct. 5-7.

[Nic01] Pierre Nicodème. Fast approximate motif statistics. *Journal of Computational Biology*, 8(3):235–248, 2001.

[NSF01] Pierre Nicodème, Bruno Salvy, and Philippe Flajolet. Motif statistics. *Theoretical Computer Science*, 2001. To appear. Extended version of an article published in the proceedings of 7th Annual European Symposium on Algorithms ESA'99, Prague, July 1999.

[PBM89] P. A. Pevzner, M. Y. Borodovski, and A. A. Mironov. Linguistic of nucleotide sequences: The significance of deviation from mean statistical characteristics and prediction of the frequencies of occurrence of words. *J. Biomol. Struct. Dyn*, 6:1013–1026, 1989.

[PRdT95] B. Prum, F. Rodolphe, and E. de Turckheim. Finding words with unexpected frequencies in deoxyribonucleic acid sequences. *J. R. statist. Soc. B*, 57(1):205–220, 1995.

[Rég98] M. Régnier. A unified approach to words statistics. In *Second Annual International Conference on Computational Molecular Biology*, pages 207–213. ACM Press, New-York, 1998.

[Rég00] M. Régnier. A unified approach to word occurrences probabilities. *Discrete Applied Mathematics*, 104(1):259–280, 2000. Special issue on Computational Biology.

[RP86] E. Ruoslahti and M.D. Pierschbacher. Arg-Gly-Asp: a versatile cell recognition signal. *Cell*, 44:517–518, 1986.

[RS98a] M. Régnier and W. Szpankowski. On Pattern Frequency Occurrences in a Markovian Sequence. *Algorithmica*, 22(4):631–649, 1998.

[RS98b] Gesine Reinert and Sophie Schbath. Compound Poisson Approximations for Occurrences of Multiple Words in Markov Chains. *J. Comp. Biol.*, 5(2):223–253, 1998.

[SCD+00] J. Schultz, R.R. Copley, T. Doerks, Ponting C.P., and P. Bork. SMART: A Web-based tool for the study of genetically mobile domains. *Nucleic Acids Res.*, 28:231–234, 2000.

[SPdT95] S. Schbath, B. Prum, and É. de Turckheim. Exceptional motifs in different Markov chain models for a statistical analysis of DNA sequences. *J. Comp. Biol.*, 2(3):417–437, 1995.

[ST00] Saurabh Sinha and Martin Tompa. A Statistical Method for Finding Transcription Factor Binding Sites. In *Eight International Conference on Intelligent Systems for Molecular Biology*, pages 354–354. AAAI Press, 2000.

APPENDIX

- ***Bacteria***
  - ***Aquificales*** - *Aquifex aeolicus*
  - ***Firmicutes***
    - ***Bacillus/Clostridium group***
      - ***Bacillaceae*** - *Bacillus subtilis - Bacillus halodurans C-125*
      - ***Mycoplasmataceae***
        - ***Mycoplasma*** - *Mycoplasma genitalium - Mycoplasma pneumoniae*
        - ***Ureaplasma*** - *Ureaplasma urealyticum*
      - ***Streptococcaceae*** - *Lactococcus lactis*
    - ***Actinobacteria*** - *Mycobacterium tuberculosis H37Rv - Mycobacterium tuberculosis CDC1551 - Mycobacterium leprae*
  - ***Spirochaetales***
    - ***Spirochaetaceae***
      - ***Borrelia*** - *Borrelia burgdorferi*
      - ***Treponema*** - *Treponema pallidum subsp. pallidum*
  - ***Thermotogales*** - *Thermotoga maritima*
  - ***Thermus/Deinococcus group*** - *Deinococcus radiodurans*
  - ***Planctomyces/Chlamydia/Verrucomicrobium group*** - *Chlamydia trachomatis - Chlamydia muridarum - Chlamydophila pneumoniae CWL029 - Chlamydophila pneumoniae AR39 - Chlamydophila pneumoniae J138*
  - ***Proteobacteria***
    - ***alpha subdivision***
      - ***Rickettsiales*** - *Rickettsia prowazekii* Madrid E
    - ***beta subdivision***
      - ***Neisseria meningitidis*** - *Neisseria meningitidis Z2491 (serogroup A) - Neisseria meningitidis MC58 (serogroup B)*
    - ***gamma subdivision***
      - ***Enterobacteriaceae group***
        - ***Enterobacteriaceae***
          - ***Escherichia*** - *Escherichia coli - Escherichia coli O157:H7 - Escherichia coli O157:H7 EDL933*
        - ***Buchnera*** - *Buchnera sp. APS*
      - ***Pasteurellaceae*** - *Haemophilus influenzae Rd - Pasteurella multocida*
      - ***Pseudomonadaceae*** - *Pseudomonas aeruginosa*
      - ***Xanthomonas group*** - *Xylella fastidiosa*
      - ***Vibrionaceae*** - *Vibrio cholerae*
    - ***delta/epsilon subdivisions*** - *Campylobacter jejuni - Helicobacter pylori 26695 - Helicobacter pylori J99*
  - ***Cyanobacteria***

    Synechocystis PCC6803 - Nostoc sp. PCC7120

Revised January, 23, 2002

`http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/new_micr.html`

Figure 4: Extract from Microbial Genomes Taxonomy at NCBI.

	PS00010		PS00016		PS00027		PS00028		PS00039		PS00043		PS00044	
	μ_n	ω_n	μ_n	ω_n	μ_n	ω_n	μ_n	ω_n	μ_n	ω_n	μ_n	ω_n	μ_n	ω_n
dros-mela	.28	133	1187.45	883	.75	80	4.84	1419	.68	23	1.41	2	2.96	7
caen-eleg	.58	113	1124.24	987	1.33	59	5.67	412	.92	23	2.07	9	3.50	6
sacc-cere	.06	0	375.28	316	.49	6	.72	82	.40	20	.74	0	1.10	2
arab-thal	.50	11	2104.73	1738	2.04	58	6.21	236	1.47	28	3.37	2	5.69	8
aero-pern	.00	0	163.23	184	.16	0	.07	2	.06	1	.72	2	1.23	2
arch-fulg	.01	0	134.95	125	.16	0	.07	1	.16	2	.54	1	.61	2
meth-jann	.01	0	67.31	84	.09	0	.06	0	.13	1	.31	0	.16	1
meth-ther	.01	0	168.11	173	.09	0	.09	0	.14	2	.39	0	.49	0
pyro-abys	.00	0	102.87	105	.15	0	.01	0	.11	0	.52	1	.40	0
pyro-hori	.00	0	92.28	96	.16	0	.02	1	.09	1	.55	0	.45	0
halo-nrc1	.00	0	309.74	292	.10	0	.05	4	.28	0	.40	0	1.23	2
ther-acid	.00	0	103.86	68	.10	1	.02	0	.09	2	.32	0	.40	3
aqui-aeol	.00	0	69.83	74	.09	0	.03	2	.09	0	.33	0	.24	4
baci-halo	.00	0	200.05	170	.27	1	.12	1	.22	3	.77	8	1.07	5
baci-subt	.00	0	180.22	102	.26	0	.13	1	.22	3	.71	11	.99	21
myco-geni	.00	0	12.43	8	.05	0	.01	0	.02	0	.08	0	.08	0
myco-pneu	.00	0	22.81	13	.06	0	.01	0	.03	0	.10	0	.14	2
urea-urea	.00	0	14.81	18	.06	0	.01	0	.04	0	.09	0	.06	0
lact-lact	.00	0	82.98	49	.16	0	.01	0	.13	1	.41	3	.57	8
myco-lepr	.00	0	180.51	149	.16	0	.07	0	.09	2	.56	1	1.33	4
myco-tube	.00	0	557.93	409	.34	0	.15	0	.18	4	1.14	4	3.57	7
borr-burg	.00	0	28.23	26	.05	0	.01	0	.06	0	.16	0	.10	0
trep-pall	.01	0	81.88	77	.08	0	.31	1	.04	1	.25	2	.55	1
ther-mari	.00	0	110.88	102	.14	0	.03	1	.12	0	.45	1	.40	0
dein-radi	.00	0	327.35	247	.20	0	.05	0	.12	1	.74	4	2.04	3
chla-muri	.01	0	44.33	34	.07	0	.15	0	.04	0	.19	0	.28	0
chla-pneu	.01	0	46.59	45	.07	0	.17	0	.04	0	.21	2	.28	0
chla-trac	.01	0	43.40	33	.06	0	.15	0	.04	0	.19	0	.28	1
rick-prow	.00	0	24.79	25	.07	0	.04	0	.04	0	.20	0	.16	0
neis-mena	.00	0	129.41	91	.13	0	.09	1	.10	2	.33	2	.64	6
neis-menb	.00	0	129.13	87	.13	0	.10	1	.10	2	.31	2	.62	6
esch-coli	.01	0	288.99	228	.45	1	.31	1	.20	6	.95	17	1.73	45
buch-aphi	.00	0	16.72	18	.05	0	.04	0	.02	1	.09	0	.06	1
haem-infl	.00	0	78.23	69	.14	0	.08	0	.09	3	.32	3	.46	8
past-mult	.01	0	95.17	82	.19	0	.15	3	.10	3	.45	2	.65	9
pseu-aeru	.01	0	634.32	482	.47	0	.27	2	.32	6	1.29	13	2.81	89
xyla-fast	.01	0	200.56	170	.22	0	.25	1	.09	3	.54	1	1.11	6
vibr-chol	.01	1	191.57	153	.34	1	.24	0	.17	10	.76	4	1.33	36
camp-jeju	.01	0	44.69	39	.09	0	.07	1	.11	0	.23	0	.22	1
heli-2pyl	.01	0	46.81	45	.08	0	.09	0	.08	0	.22	0	.25	0
heli-Jpyl	.01	0	47.36	47	.08	0	.09	0	.08	0	.22	0	.25	0
syne-cyan	.01	1	192.26	190	.30	0	.12	1	.13	2	.67	4	1.05	6

PS00010 Aspartic acid and asparagine hydroxylation site
PS00016 Cell attachment sequence
PS00027 'Homeobox' domain signature
PS00028 Zinc finger, C2H2 type
PS00039 AD-box subfamily ATP-dependent helicases
PS00043 Bacterial regulatory proteins, gntR family
PS00044 Bacterial regulatory proteins, lysR family

Table 3: Unscaled expectations μ_n and observations ω_n for the characteristic motifs.

	PS00010					PS00016					
	μ_n	ω_n	μ_s	ω_s	γ_s	μ_n	ω_n	μ_s	ω_s	γ_s	$n/10^6$
dros-mela	.28	133	.04	20.10	14.79	1187.45	883	179.43	133.43	-3.57	6.62
caen-eleg	.58	113	.07	14.37	11.41	1124.24	987	142.98	125.52	-1.15	7.86
sacc-cere	.06	0	.02	0.00	-.03	375.28	316	127.77	107.59	-1.53	2.94
arab-thal	.50	11	.04	.97	.09	2104.73	1738	186.19	153.75	-2.43	11.30
aero-pern	.00	0	.00	0.00	-.00	163.23	184	255.52	288.04	1.78	.64
arch-fulg	.01	0	.01	0.00	-.01	134.95	125	204.21	189.16	-.68	.66
meth-jann	.01	0	.02	0.00	-.02	67.31	84	134.00	167.21	2.77	.50
meth-ther	.01	0	.01	0.00	-.01	168.11	173	320.10	329.42	.20	.53
pyro-abys	.00	0	.00	0.00	-.00	102.87	105	192.00	195.97	.25	.54
pyro-hori	.00	0	.00	0.00	-.00	92.28	96	162.30	168.84	.00	.57
halo-nrc1	.00	0	.00	0.00	-.00	309.74	292	496.37	467.94	-1.07	.62
ther-acid	.00	0	.00	0.00	-.00	103.86	68	229.22	150.07	-5.54	.45
aqui-aeol	.00	0	.00	0.00	-.00	69.83	74	142.96	151.49	.26	.49
baci-halo	.00	0	.00	0.00	-.00	200.05	170	171.10	145.40	-1.72	1.17
baci-subt	.00	0	.00	0.00	-.00	180.22	102	147.90	83.71	-5.74	1.22
myco-geni	.00	0	.01	0.00	-.01	12.43	8	70.71	45.52	-3.18	.18
myco-pneu	.00	0	.00	0.00	-.01	22.81	13	95.86	54.64	-4.54	.24
urea-urea	.00	0	.00	0.00	-.01	14.81	18	65.04	79.05	1.40	.23
lact-lact	.00	0	.00	0.00	-.00	82.98	49	126.50	74.70	-4.97	.66
myco-lepr	.00	0	.00	0.00	-.00	180.51	149	347.09	286.51	-3.36	.52
myco-tube	.00	0	.00	0.00	-.00	557.93	409	423.57	310.51	-5.79	1.32
borr-burg	.00	0	.01	0.00	-.01	28.23	26	80.06	73.73	-.61	.35
trep-pall	.01	0	.04	0.00	-.05	81.88	77	234.00	220.05	-.52	.35
ther-mari	.00	0	.00	0.00	-.00	110.88	102	190.50	175.25	-.74	.58
dein-radi	.00	0	.00	0.00	-.00	327.35	247	344.05	259.60	-4.77	.95
chla-muri	.01	0	.03	0.00	-.03	44.33	34	136.64	104.80	-2.82	.32
chla-pneu	.01	0	.03	0.00	-.03	46.59	45	128.80	124.41	-.28	.36
chla-trac	.01	0	.03	0.00	-.03	43.40	33	139.05	105.73	-2.93	.31
rick-prow	.00	0	.01	0.00	-.02	24.79	25	88.86	89.59	.09	.28
neis-mena	.00	0	.01	0.00	-.01	129.41	91	222.32	156.33	-4.65	.58
neis-menb	.00	0	.01	0.00	-.01	129.13	87	225.02	151.60	-5.20	.57
esch-coli	.01	0	.01	0.00	-.01	288.99	228	210.36	165.96	-3.19	1.37
buch-aphi	.00	0	.02	0.00	-.02	16.72	18	89.40	96.24	.76	.19
haem-infl	.00	0	.01	0.00	-.01	78.23	69	148.51	130.98	-1.17	.53
past-mult	.01	0	.01	0.00	-.01	95.17	82	142.54	122.81	-1.40	.67
pseu-aeru	.01	0	.00	0.00	-.01	634.32	482	341.63	259.59	-4.65	1.86
xyla-fast	.01	0	.01	0.00	-.01	200.56	170	269.41	228.36	-2.54	.74
vibr-chol	.01	1	.01	.86	.18	191.57	153	165.70	132.34	-2.63	1.16
camp-jeju	.01	0	.02	0.00	-.03	44.69	39	88.76	77.46	-.82	.50
heli-2pyl	.01	0	.01	0.00	-.02	46.81	45	95.19	91.51	-.26	.49
heli-Jpyl	.01	0	.01	0.00	-.02	47.36	47	96.05	95.33	-.64	.49
syne-cyan	.01	1	.01	.97	.04	192.26	190	187.20	185.00	-.04	1.03

Table 4: From unscaled expectations μ_n and observations ω_n to scaled expectations μ_s and observations ω_s and to scaled calibrations γ_s for the motifs PS00010 and PS00016.