# Genome Comparison based on Motif Statistics

*Pierre Nicodème*

Laboratory Statistics and Genome
La Génopole - Evry

and

Algorithm Project
INRIA - Rocquencourt

`nicodeme@genopole.cnrs.fr`

# Motivation : Prosite Motifs

```
AC   PS00723;
DE   Polyprenyl synthetases signature 1.
...
PA   [LIVM](2)-x-D-D-x(2,4)-D-x(4)-R-R-[GH].

...
DR   P14324, FPPS_HUMAN, T; ... P49353, FPPS_MAIZE, T;
DR   P08524, FPPS_YEAST, T; ... P08836, FPPS_CHICK, P;
...
```

Biological pertinence of motifs

with respect to target proteoms

# Content of the talk

1. Exact method
2. Approximate method
   (a) Generating function for non-self-overlapping motifs
   (b) Asymptotic Poisson distribution for rare occurrences
3. Biological application
   – Characteristic motifs in Eukaryotes, Archae-Bacteria and Bacteria

## Regular expressions

$$R = (a \cdot b \cdot a + (c^4 \cdot e)^\star \cdot b \cdot b)^\star$$

### Operators

| | |
|---|---|
| $+$ | Union |
| $\cdot$ | Concatenation |
| $\star$ | Star-operator $(A^\star = \epsilon + A + A^2 + A^3 + \ldots)$ |

## Prosite motifs

```
PA = [LIVM](2)-x-D-D-x(2,4)-D-x(4)-R-R-[GH].
```

$$PA = (L+I+V+M)^2 \cdot \Sigma \cdot D \cdot D \cdot (\Sigma^2 + \Sigma^3 + \Sigma^4) \cdot D \cdot \Sigma^4 \cdot R \cdot R \cdot (G+H)$$

$$\text{with } \Sigma = \begin{array}{l} A+R+N+D+C+Q+E+G+H+I \\ +L+K+M+F+P+S+T+W+Y+V \end{array}$$

$$\Sigma^k = \Sigma \cdot \Sigma \cdot \quad \cdots \quad \cdot \Sigma \qquad (k \text{ times})$$

# Main counting tool

$R$ given regular expression.

$X_n$ number of occurrences of words of $R$ in a text of length $n$.

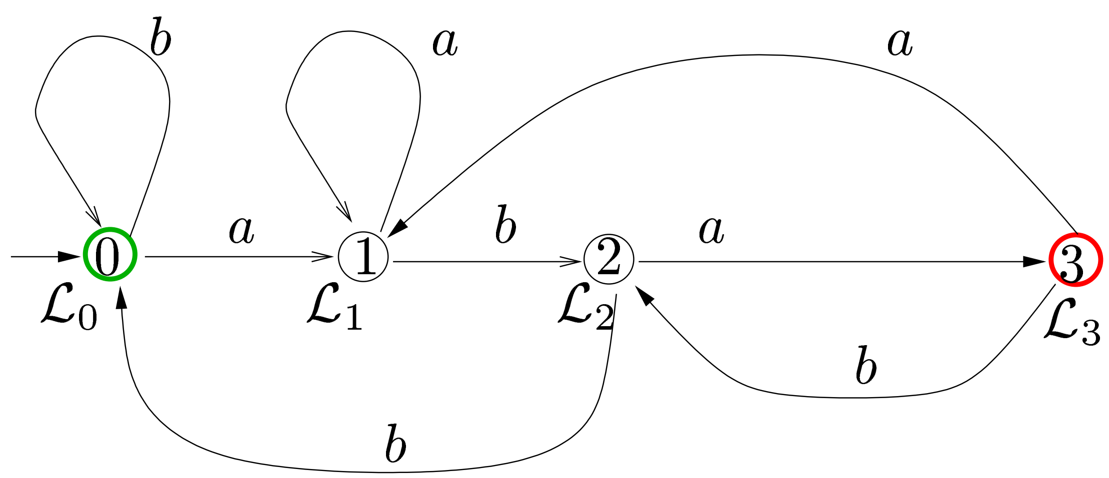
$$F(z,u) = \sum_{n,k} \Pr(X_n = k) u^k z^n = \sum_{n,k} f_{n,k} u^k z^n.$$

# Exact method

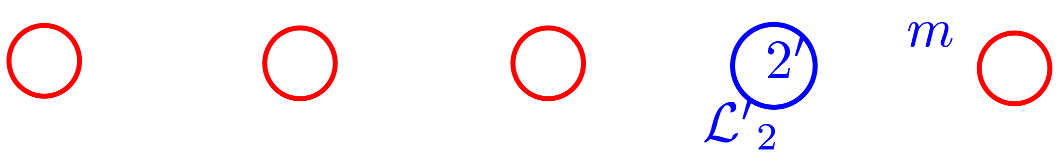
Automaton Recognizing $\Sigma^* R$

$$\Sigma = \{a, b\} \quad R = aba, \quad \Sigma^* R = \Sigma^* aba$$

$$aabbaba\bullet bbabbaba\bullet aaaba\bullet bbbb$$



marking ($\Sigma' = \Sigma \cup \{m\}$)




$aaaabambamaaabamaa$

Chomsky-Schützenberger

$$\mathcal{L}_0 = a\mathcal{L}_1 + b\mathcal{L}_0 + \epsilon, \qquad \ldots,$$

$$\mathcal{L}_2 = a\mathcal{L}'_2 + b\mathcal{L}_0 + \epsilon, \qquad \mathcal{L}'_2 = m\mathcal{L}_3,$$

$$L(a,b,m) = \frac{1 + ab(1-m)}{1 - a - b + ab(1-m) - ab^2(1-m)},$$

$$F(z,u) = L(\omega_a z, \omega_b z, u).$$

# Experimental observations

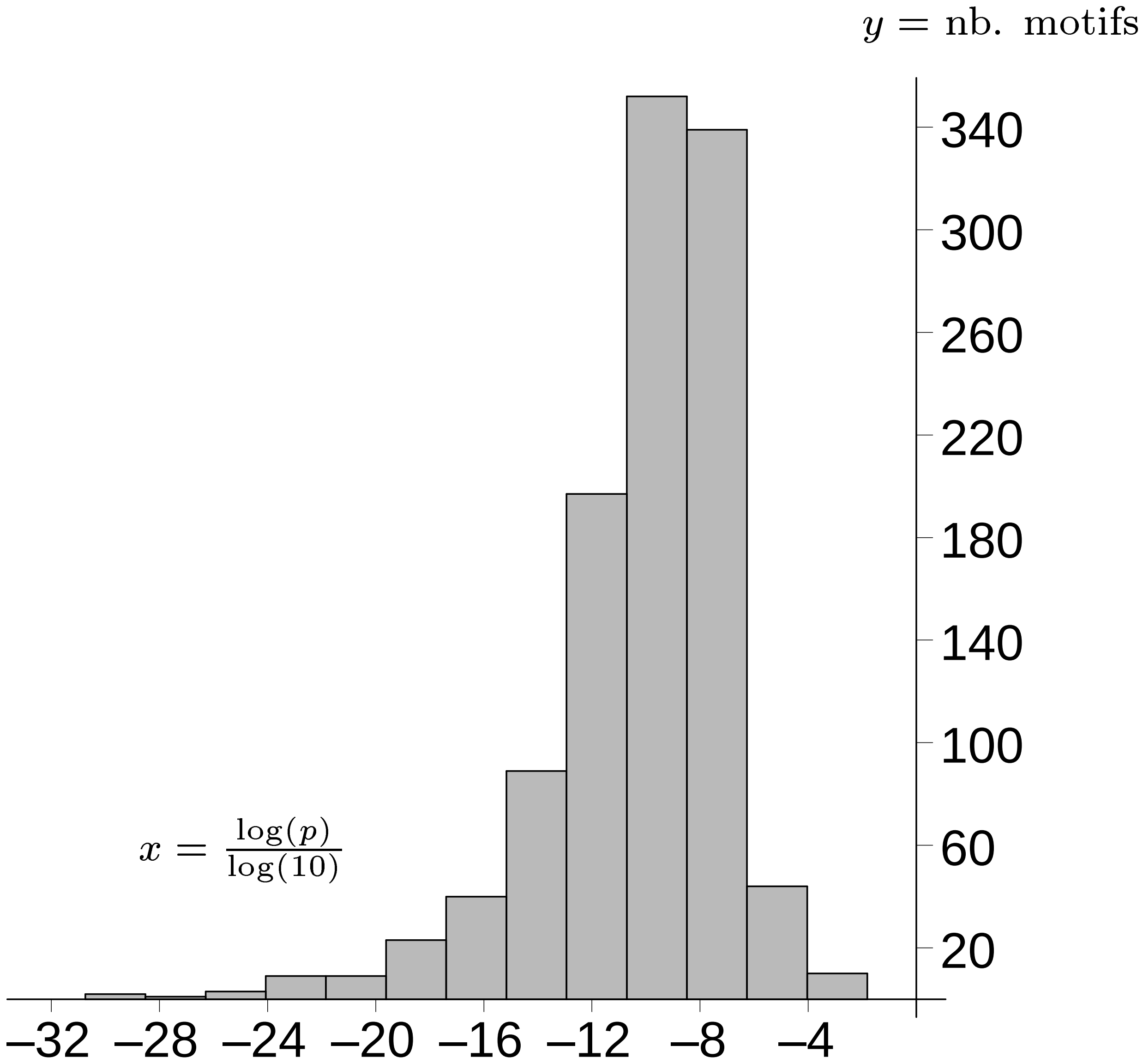



Figure 1: Histogram of probability of occurrence $p$ of a motif at a given position for the 1118 motifs exactly computed. (ProDom distribution for amino-acids)

# Approximate method

Matches for non-self-overlapping languages

$R$ non-self-overlapping, O.G.F. $R(z)$

$F(z,u) = \sum f_{n,k} u^k z^n$, O.G.F. nb. matches

$\mu_n, \sigma_n^2$ asympt. expect. and var. for nb. matches

Theorem 1:

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

(b) $$\mu_n = R(1) \times n + R(1) - R'(1) + O(A^n)$$

(c) $$\sigma_n^2 = \{R(1) + R(1)^2 - 2R(1)R'(1)\} \times n + \kappa(R(1), R'(1), R''(1)) + O(A^n)$$

$$(A < 1)$$

# A non-ambiguity lemma

Warning

$A = B = C = \{a, b, ab\}$

$abab = a \cdot b \cdot ab = ab \cdot a \cdot b \in ABC$

$\Rightarrow ABC$ ambiguous

while $AB$ and $BC$ non-ambiguous

---

$R$ regular expression

$N = \Sigma^* - \Sigma^* R \Sigma^*$ lang. without occ. of $R$

Lemma:

$R$ non self-overlapping and non nested

$\Rightarrow \quad N(RN)^k$ non-ambiguous $\forall k$

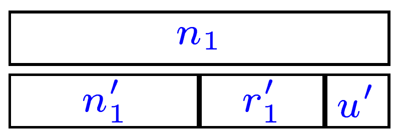
# Proof of the lemma

Let us suppose $NRN$ is ambiguous

let $w = n_1 r_1 n_2 = n'_1 r'_1 n'_2 \in NRN$, $\{n_1, n_2, n'_1, n'_2\} \subset N$, $\{r_1, r'_1\} \subset R$

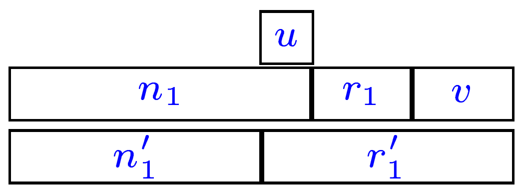
Let us suppose $|n_1| > |n'_1|$ $\quad$ ($|n_1| < |n'_1|$ is symmetrical)

$|n_1| > |n'_1 r'_1|$



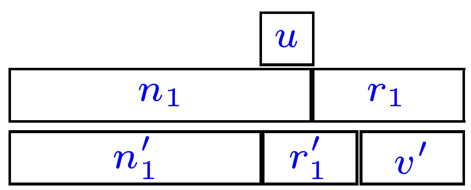

$\Rightarrow n_1 \notin N$

$|n_1 r_1| < |n'_1 r'_1|$




$\Rightarrow$ $r'_1 = u r_1 v$
$R$ nested

$|n_1 r_1| > |n'_1 r'_1|$




$\Rightarrow$ $u r_1 = r'_1 v$
$R$ overlapping

$\Rightarrow$ $n_1 = n'_1$ $\Rightarrow$ $r_1 = r'_1$ $\Rightarrow$ $n_2 = n'_2$

(elsewhere $R$ is overlapping)

$\Rightarrow$ $NRN$ non-ambiguous

$\Rightarrow$ $N(RN)^k$ $(k > 1)$ non-ambiguous (same proof)

# Proof of Theorem 1

$\phi_R$ injection $\Sigma^\star \to (\Sigma + u)^\star$

$t \to t_u = \phi_R(t)$

by inserting $u$ after each match in $t$

$\phi_R(\Sigma^\star) = \bigcup_{k\geq 0} N(RuN)^k$

$$\Rightarrow F(z,u) = \sum_{k\geq 0} N(z)(uN(z)R(z))^k = \frac{N(z)}{1 - uN(z)R(z)}$$

$$\Rightarrow F(z,1) = \frac{1}{1-z} = \frac{N(z)}{1 - N(z)R(z)}$$

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

# Radius of convergence for prefix-free languages

$\mathcal{L}$ prefix-free $\quad \Leftrightarrow \quad \forall u, v \in \mathcal{L}, \quad \not\exists w, \ u = v.w$

$\mathcal{L}_k = \mathcal{L} \cup \Sigma^k$

$j > k \quad \Rightarrow \quad \mathcal{L}_j \subset \Sigma^j - \mathcal{L}_k \Sigma^{j-k}$

fix $i$

The sets $\mathcal{M}_{i,j} = \mathcal{L}_j \Sigma^{i-j}$ $(j < i)$ are disjoints

$$\bigcup_{j \leq i} \mathcal{M}_{i,j} \subset \Sigma^i, \quad \Sigma^k(1) = 1 \ (\forall k)$$

$$\Rightarrow \quad \sum_{j \leq i} M_{i,j}(1) \leq 1 \ (\forall i), \quad M_{i,j}(1) = L_j(1)$$

$$\Rightarrow \quad L(1) = \sum_{i \geq 1} L_i(1) \leq 1$$

$\tau$ radius of convergence of $L(z)$

$L(1) = \sum l_n z^n |_{z=1}$ bounded, $l_n \geq 0 \ \Rightarrow \ \tau > 1$

# Poisson law for rare occurrences

$L$ non self-overlapping, non nested, O.G.F. $L(z)$

$p_{n,k}$ prob. $k$ matches in text size $n$

Theorem 2:

(a) $\exists \rho \in \mathbb{R}^+$, $\rho < 1$, and $1 - \rho + L(\rho) = 0$

(b) for $k = O(1)$, $\exists \kappa > \rho$

$p_{n,k} = \sum_{j=1}^{k+1} (-1)^j a_j \binom{n}{j-1} \rho^{-(n+j)} + O(\kappa^{-n})$

and

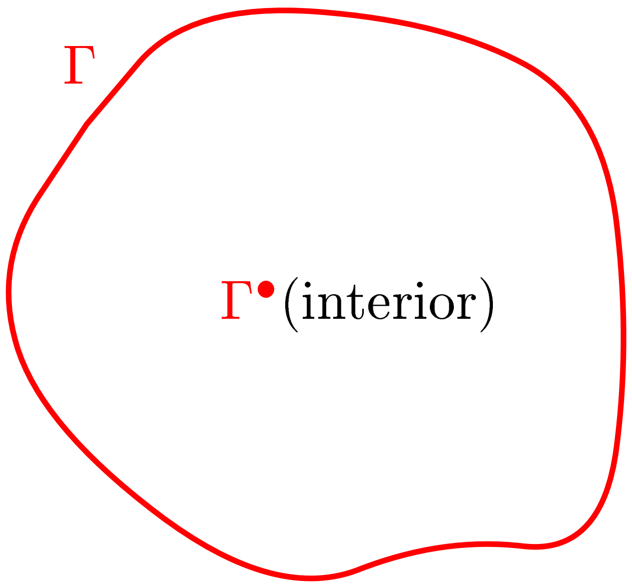
$a_{k+1} = \frac{L(\rho)^k}{\rho^{k+1}(L'(\rho)-1)^{k+1}}$

(c) asymptotically, for $k = O(1)$

$$p_{n,k} = \frac{\rho^{-n-1}}{1 - L'(\rho)} \frac{(\alpha n)^k}{k!} \left(1 + O\left(\frac{1}{n}\right)\right),$$

$$\alpha = \left(\frac{\rho^{-1} L(\rho)}{1 - L'(\rho)}\right)$$

# Rouché's Theorem




1) $f$ and $g$ differentiable in $\{\Gamma^\bullet$ - finite nb. poles$\}$

2) $0 < |f(z)| < \infty \quad$ for $z \in \Gamma$

3) $|g(z)| < |f(z)| \quad$ for $z \in \Gamma$

$\Rightarrow \quad f + g \quad$ and $\quad f$

have same number of zeroes inside $\Gamma^\bullet$

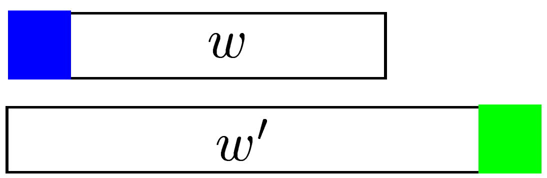
Demonstration: homotopy on

$$\frac{1}{2\pi i}\int_\Gamma \frac{f' + tg'}{f + tg}, \qquad t \in [0, 1]$$

# Proof of Theorem 2

$$F(z,u) = \frac{1}{1-z+(1-u)L(z)} = \sum_{k\geq 0} \frac{L(z)^k}{(1-z+L(z))^{k+1}} u^k$$

asymptotics come from zeroes of $1-z+L(z)$




$B = \{b, \quad b \cdot w \in L\}$

$T = \{t, \quad w' \cdot t \in L\}$

$L$ non overlapping $\quad \Rightarrow \quad T \subseteq \Sigma - B$

$b \in B, t \in T \quad \Rightarrow bbt$ and $btt \quad$ (of $B\Sigma T$) overlap

$\Rightarrow \quad L \subset B\Sigma^\star T \subseteq \Phi = B\Sigma^\star(\Sigma - B)$

$L(z) < \Phi(z) = \pi_B z \frac{1}{1-z}(1-\pi_B)z \leq \frac{z^2}{4(1-z)}$

On $\Gamma = \{|z| = 2\} \qquad |L(z)| < 1 \leq |1-z|$

Apply Rouché's Theorem with

$f = 1-z, \quad f+g = 1-z+L(z)$

$\Rightarrow 1-z+L(z)$ has only one root $\rho$ inside $\Gamma$

Perron-Frobenius $\quad \Rightarrow \quad \rho \in \mathbb{R}^+$

$L \quad$ prefix-free $\quad \Rightarrow \quad 1 < \rho < \tau$ (radius conv $(f+g)$)

# Computation of $R(z), R'(z), R(1), R'(1)$

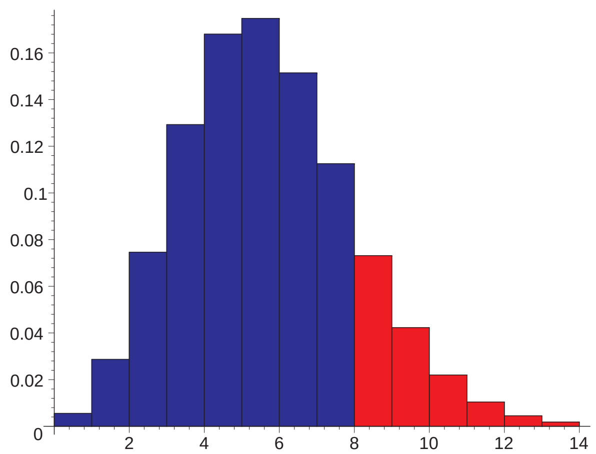
PA `[LIVM](2)-x-D-D-x(2,4)-D-x(4)-R-R-[GH]`

$R(z) = (p_L + p_I + p_V + p_M)^2 z^2 \times z \times p_D z \times p_D z \times (z^2 + z^3 + z^4) \times p_D z \times z^4 \times p_D z \times p_D z \times (p_G + p_H) z$

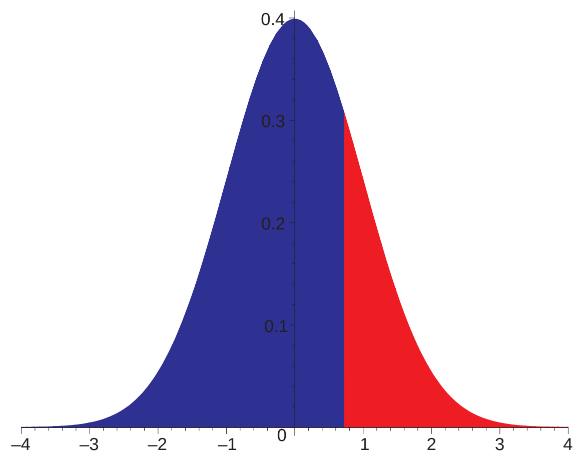
$= (p_L + p_I + p_V + p_M)^2 p_D^5 (p_G + p_H) z^{13} (z^2 + z^3 + z^4)$

$\rightarrow R'(z), R(1), R'(1)$

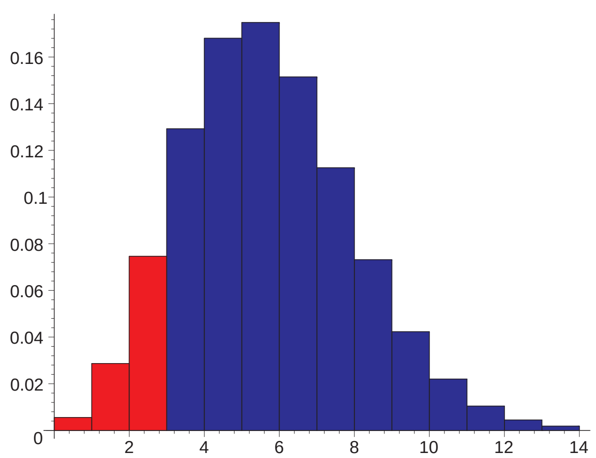
# Gaussian normalization



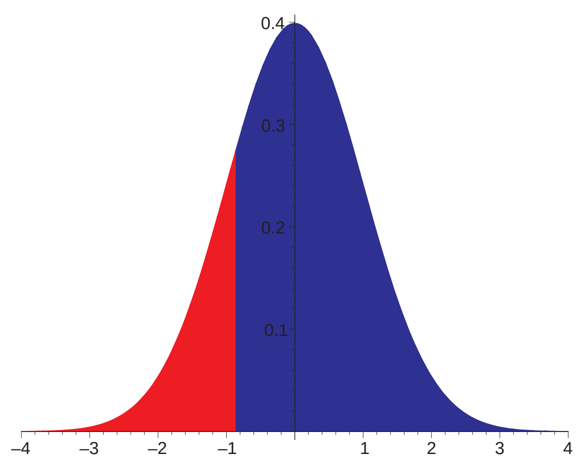

$\omega = 8$




$\gamma = +0.72$




$\omega = 2$




$\gamma = -0.81$

Poisson expectation $\mu = 5.2$

Normal distribution

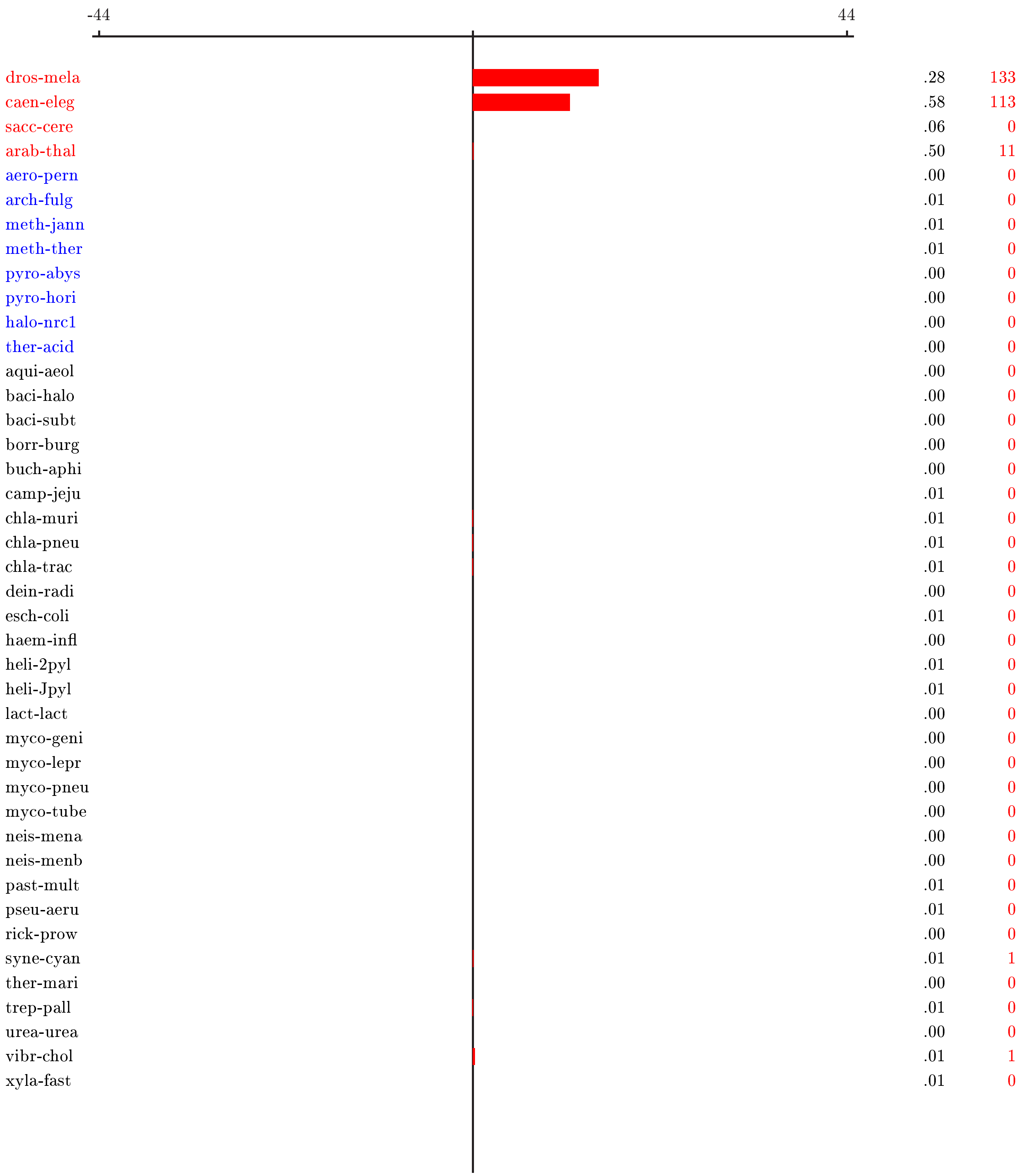
$\omega$ number of matches observed

# Comparison of 42 proteomes

| | Espèce | Longueur protéome |
|---|---|---|
| *Eukaryotes* | dros-mela | 6617803 |
| | caen-eleg | 7863058 |
| | sacc-cere | 2937205 |
| | arab-thal | 11304307 |
| *Archae-bacteria* | aero-pern | 638811 |
| | arch-fulg | 660832 |
| | meth-jann | 502362 |
| | meth-ther | 525164 |
| | pyro-abys | 535786 |
| | pyro-hori | 568584 |
| | halo-nrc1 | 624005 |
| | ther-acid | 453115 |
| *Bacteria* | aqui-aeol | 488484 |
| | baci-halo | 1169204 |
| | baci-subt | 1218487 |
| | borr-burg | 352644 |
| | buch-aphi | 187023 |
| | camp-jeju | 503501 |
| | chla-muri | 324428 |
| | chla-pneu | 361707 |
| | chla-trac | 312118 |
| | dein-radi | 951455 |
| | esch-coli | 1373785 |
| | haem-infl | 526801 |
| | heli-2pyl | 491768 |
| | heli-Jpyl | 493049 |
| | lact-lact | 655989 |
| | myco-geni | 175729 |
| | myco-lepr | 520057 |
| | myco-pneu | 237930 |
| | myco-tube | 1317198 |
| | neis-mena | 582084 |
| | neis-menb | 573863 |
| | past-mult | 667675 |
| | pseu-aeru | 1856757 |
| | rick-prow | 279044 |
| | syne-cyan | 1027015 |
| | ther-mari | 582037 |
| | trep-pall | 349913 |
| | urea-urea | 227717 |
| | vibr-chol | 1156096 |
| | xyla-fast | 744437 |

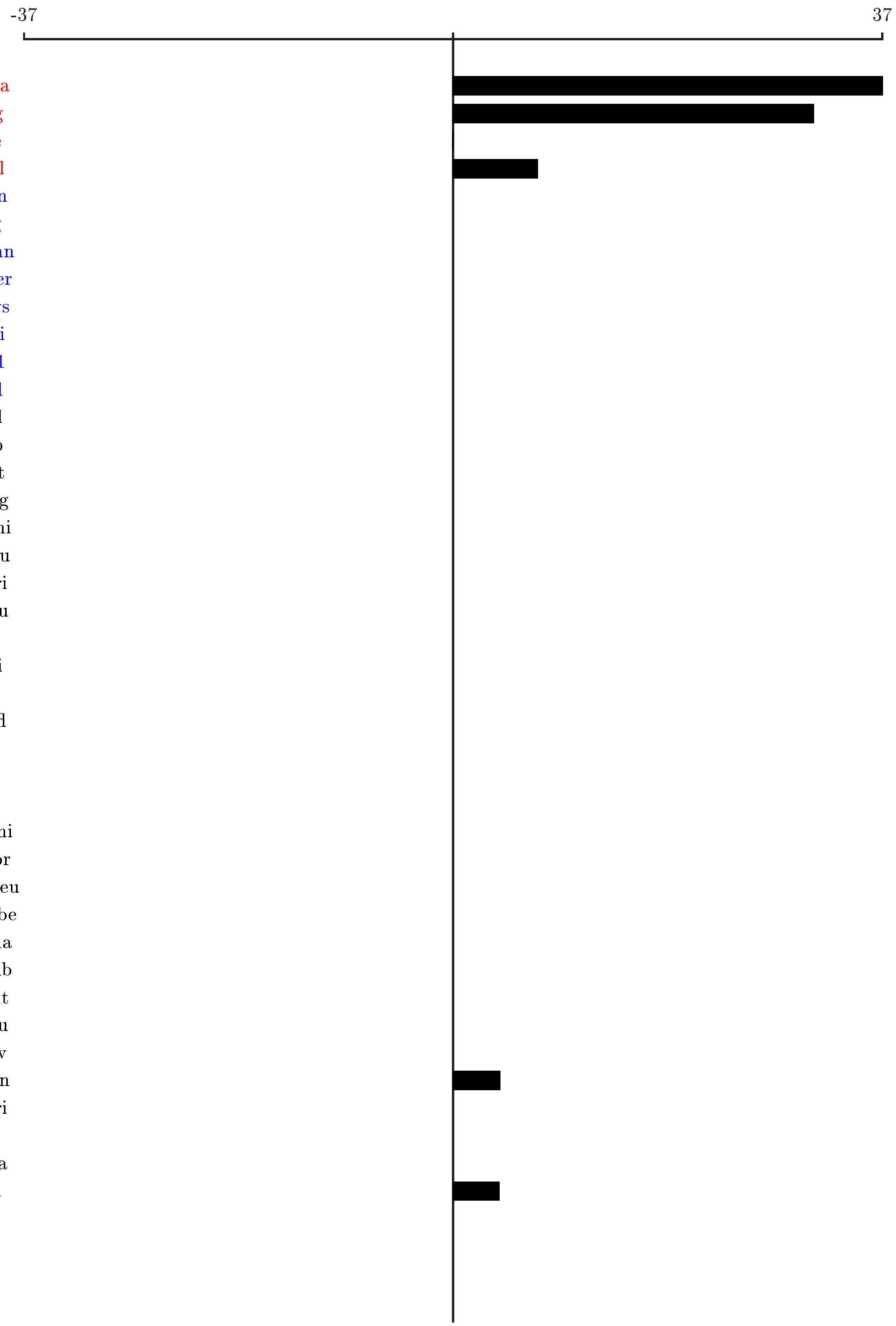
PS00010 - ASX-HYDROXYL;PATTERN. Aspartic acid and asparagine hydroxylation site.

C-x-[DN]-x(4)-[FY]-x-C-x-C



PS00010 ASX-HYDROXYL;PATTERN. Aspartic acid and asparagine hydroxylation site.

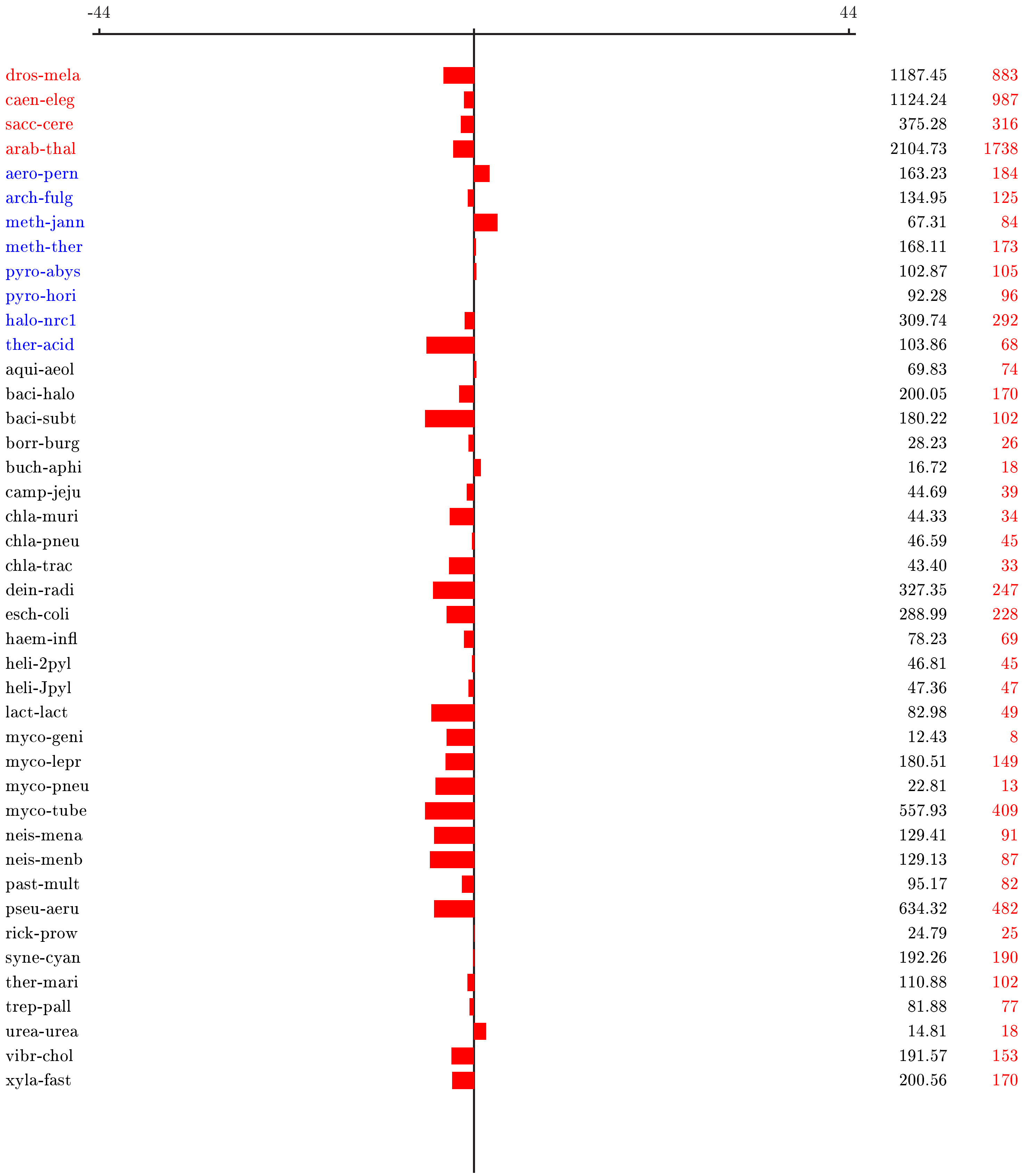
C-x-[DN]-x(4)-[FY]-x-C-x-C




| | | |
|---|---|---|
| dros-mela | .28 | 133 |
| caen-eleg | .58 | 113 |
| sacc-cere | .06 | 0 |
| arab-thal | .50 | 11 |
| aero-pern | .00 | 0 |
| arch-fulg | .01 | 0 |
| meth-jann | .01 | 0 |
| meth-ther | .01 | 0 |
| pyro-abys | .00 | 0 |
| pyro-hori | .00 | 0 |
| halo-nrc1 | .00 | 0 |
| ther-acid | .00 | 0 |
| aqui-aeol | .00 | 0 |
| baci-halo | .00 | 0 |
| baci-subt | .00 | 0 |
| borr-burg | .00 | 0 |
| buch-aphi | .00 | 0 |
| camp-jeju | .01 | 0 |
| chla-muri | .01 | 0 |
| chla-pneu | .01 | 0 |
| chla-trac | .01 | 0 |
| dein-radi | .00 | 0 |
| esch-coli | .01 | 0 |
| haem-infl | .00 | 0 |
| heli-2pyl | .01 | 0 |
| heli-Jpyl | .01 | 0 |
| lact-lact | .00 | 0 |
| myco-geni | .00 | 0 |
| myco-lepr | .00 | 0 |
| myco-pneu | .00 | 0 |
| myco-tube | .00 | 0 |
| neis-mena | .00 | 0 |
| neis-menb | .00 | 0 |
| past-mult | .01 | 0 |
| pseu-aeru | .01 | 0 |
| rick-prow | .00 | 0 |
| syne-cyan | .01 | 1 |
| ther-mari | .00 | 0 |
| trep-pall | .01 | 0 |
| urea-urea | .00 | 0 |
| vibr-chol | .01 | 1 |
| xyla-fast | .01 | 0 |

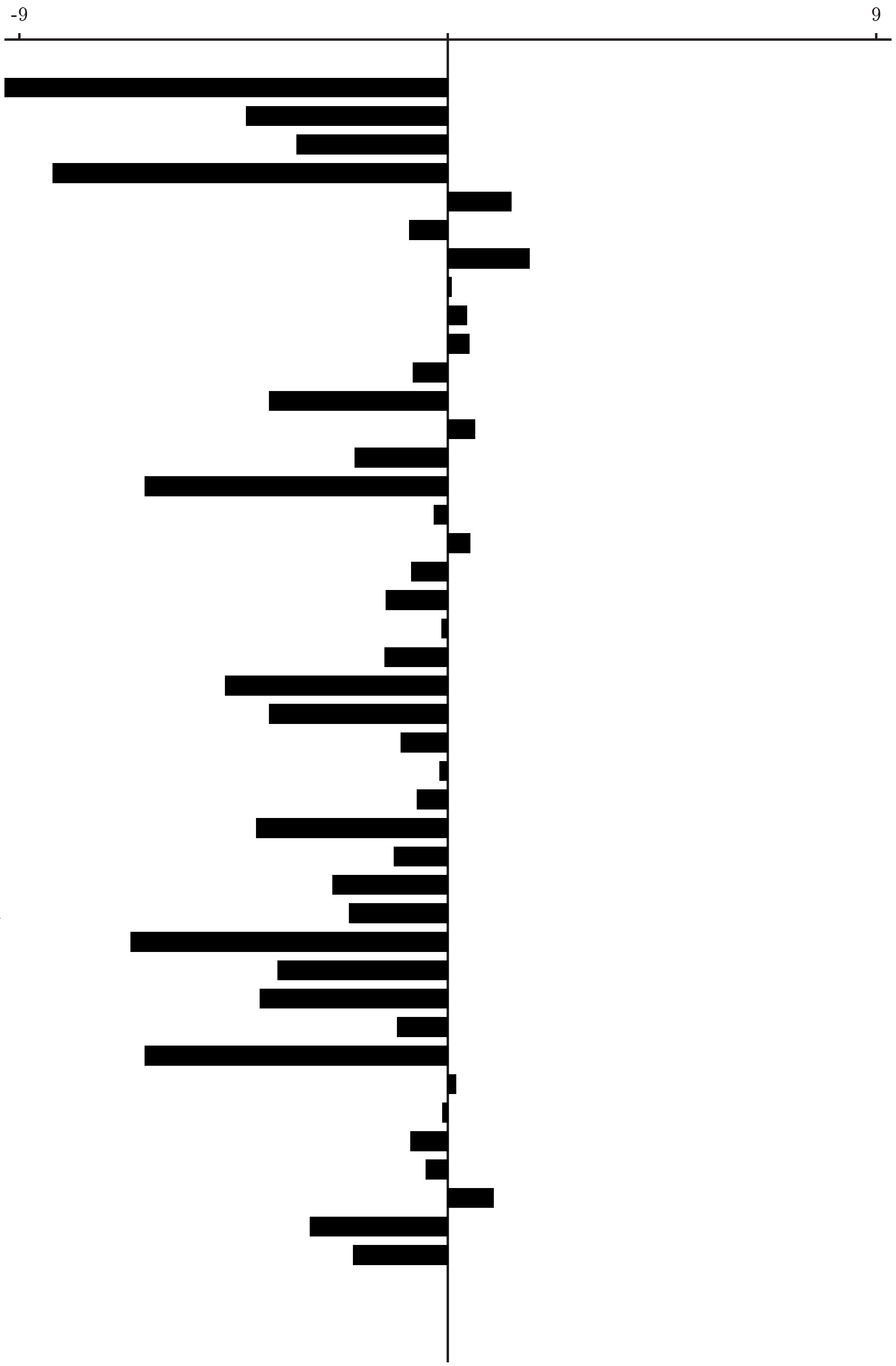
PS00016 - RGD;PATTERN. Cell attachment sequence.

R-G-D



PS00016 RGD;PATTERN. Cell attachment sequence.

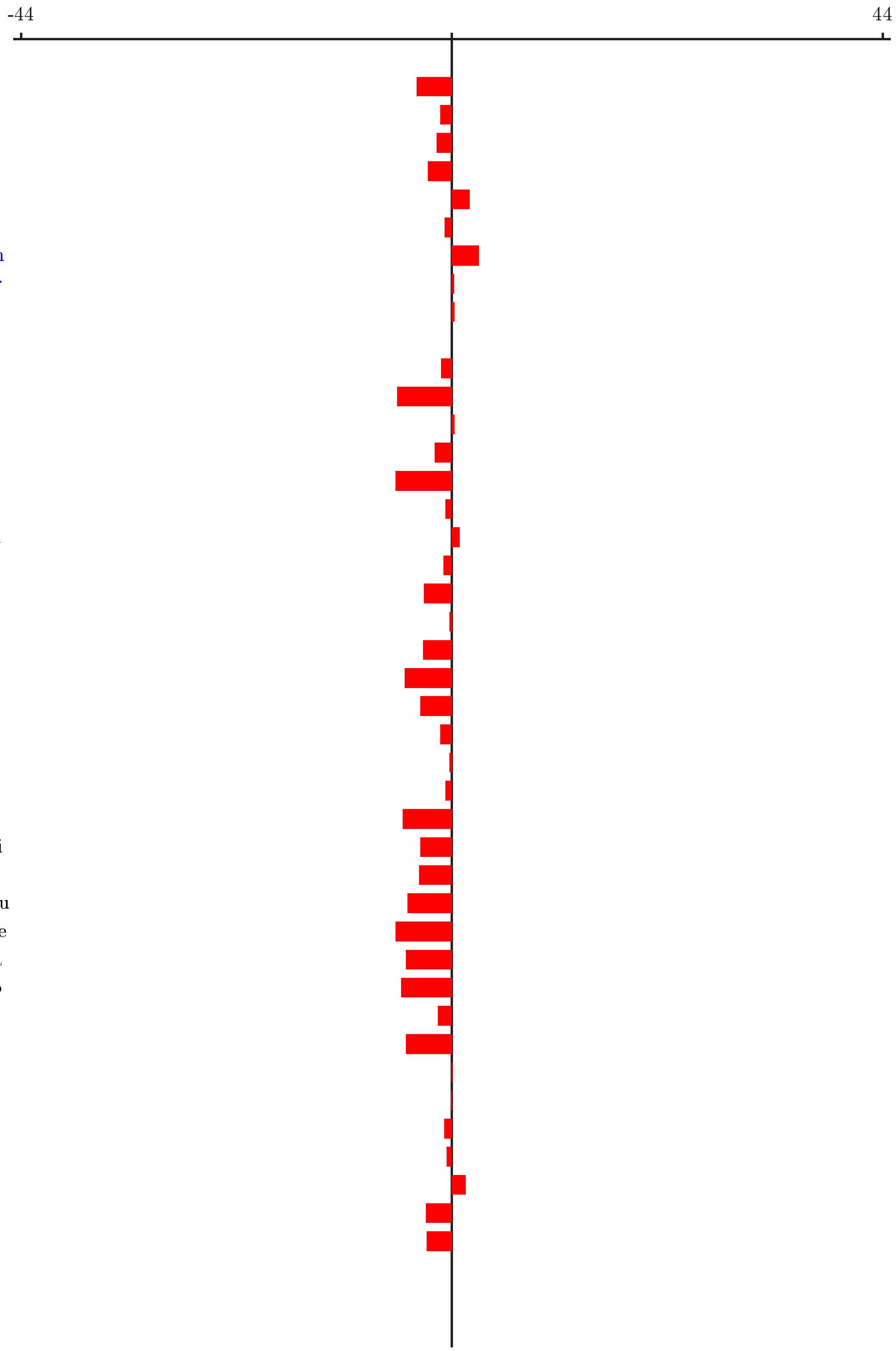
R-G-D




| | | |
|---|---|---|
| dros-mela | 1187.45 | 883 |
| caen-eleg | 1124.24 | 987 |
| sacc-cere | 375.28 | 316 |
| arab-thal | 2104.73 | 1738 |
| aero-pern | 163.23 | 184 |
| arch-fulg | 134.95 | 125 |
| meth-jann | 67.31 | 84 |
| meth-ther | 168.11 | 173 |
| pyro-abys | 102.87 | 105 |
| pyro-hori | 92.28 | 96 |
| halo-nrc1 | 309.74 | 292 |
| ther-acid | 103.86 | 68 |
| aqui-aeol | 69.83 | 74 |
| baci-halo | 200.05 | 170 |
| baci-subt | 180.22 | 102 |
| borr-burg | 28.23 | 26 |
| buch-aphi | 16.72 | 18 |
| camp-jeju | 44.69 | 39 |
| chla-muri | 44.33 | 34 |
| chla-pneu | 46.59 | 45 |
| chla-trac | 43.40 | 33 |
| dein-radi | 327.35 | 247 |
| esch-coli | 288.99 | 228 |
| haem-infl | 78.23 | 69 |
| heli-2pyl | 46.81 | 45 |
| heli-Jpyl | 47.36 | 47 |
| lact-lact | 82.98 | 49 |
| myco-geni | 12.43 | 8 |
| myco-lepr | 180.51 | 149 |
| myco-pneu | 22.81 | 13 |
| myco-tube | 557.93 | 409 |
| neis-mena | 129.41 | 91 |
| neis-menb | 129.13 | 87 |
| past-mult | 95.17 | 82 |
| pseu-aeru | 634.32 | 482 |
| rick-prow | 24.79 | 25 |
| syne-cyan | 192.26 | 190 |
| ther-mari | 110.88 | 102 |
| trep-pall | 81.88 | 77 |
| urea-urea | 14.81 | 18 |
| vibr-chol | 191.57 | 153 |
| xyla-fast | 200.56 | 170 |

PS00016 - RGD;PATTERN. Cell attachment sequence.

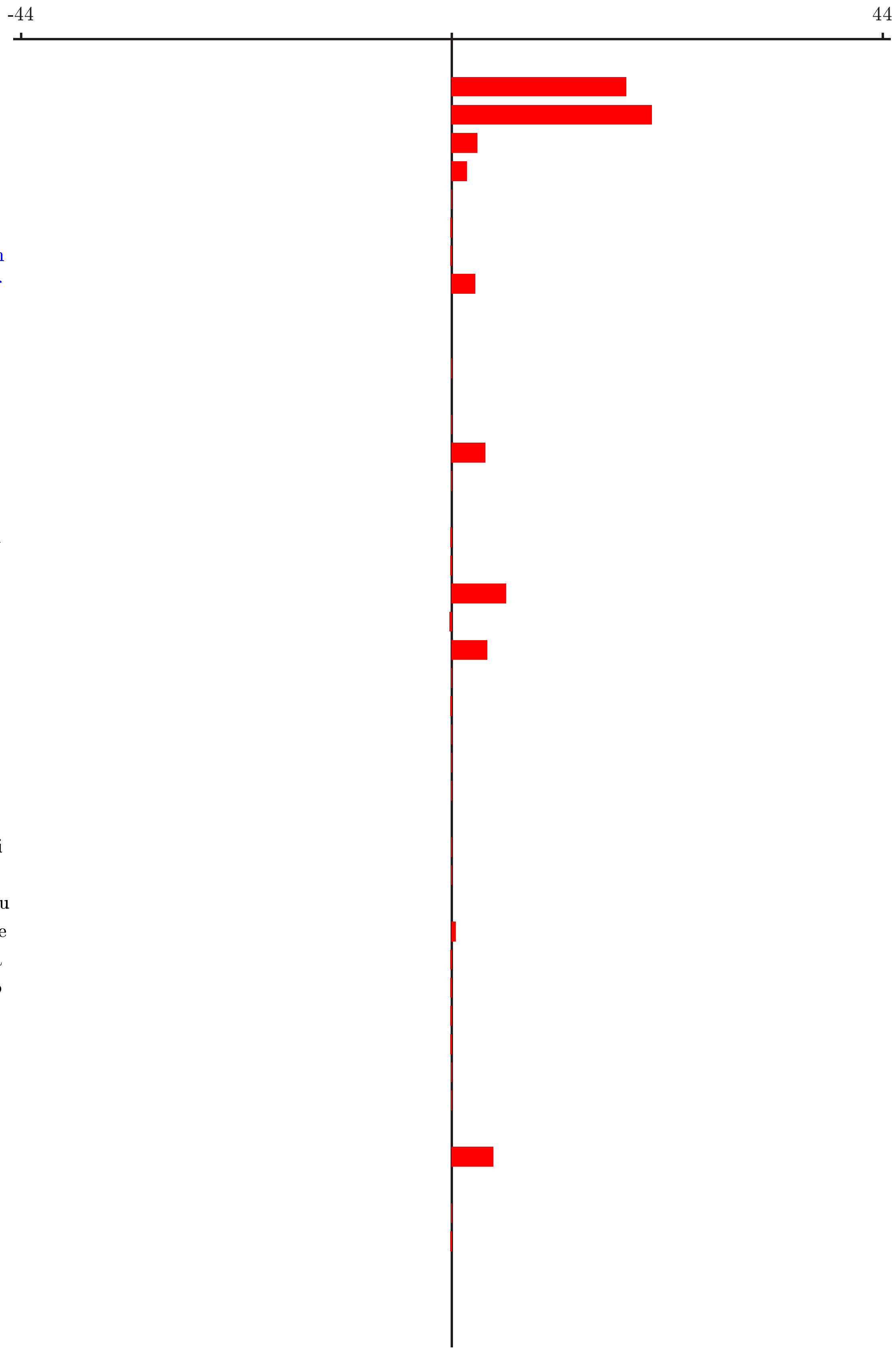
R-G-D




| | | |
|---|---|---|
| dros-mela | 1187.45 | 883 |
| caen-eleg | 1124.24 | 987 |
| sacc-cere | 375.28 | 316 |
| arab-thal | 2104.73 | 1738 |
| aero-pern | 163.23 | 184 |
| arch-fulg | 134.95 | 125 |
| meth-jann | 67.31 | 84 |
| meth-ther | 168.11 | 173 |
| pyro-abys | 102.87 | 105 |
| pyro-hori | 92.28 | 96 |
| halo-nrc1 | 309.74 | 292 |
| ther-acid | 103.86 | 68 |
| aqui-aeol | 69.83 | 74 |
| baci-halo | 200.05 | 170 |
| baci-subt | 180.22 | 102 |
| borr-burg | 28.23 | 26 |
| buch-aphi | 16.72 | 18 |
| camp-jeju | 44.69 | 39 |
| chla-muri | 44.33 | 34 |
| chla-pneu | 46.59 | 45 |
| chla-trac | 43.40 | 33 |
| dein-radi | 327.35 | 247 |
| esch-coli | 288.99 | 228 |
| haem-infl | 78.23 | 69 |
| heli-2pyl | 46.81 | 45 |
| heli-Jpyl | 47.36 | 47 |
| lact-lact | 82.98 | 49 |
| myco-geni | 12.43 | 8 |
| myco-lepr | 180.51 | 149 |
| myco-pneu | 22.81 | 13 |
| myco-tube | 557.93 | 409 |
| neis-mena | 129.41 | 91 |
| neis-menb | 129.13 | 87 |
| past-mult | 95.17 | 82 |
| pseu-aeru | 634.32 | 482 |
| rick-prow | 24.79 | 25 |
| syne-cyan | 192.26 | 190 |
| ther-mari | 110.88 | 102 |
| trep-pall | 81.88 | 77 |
| urea-urea | 14.81 | 18 |
| vibr-chol | 191.57 | 153 |
| xyla-fast | 200.56 | 170 |

PS00022 - EGF-1;PATTERN. EGF-like domain signature 1.

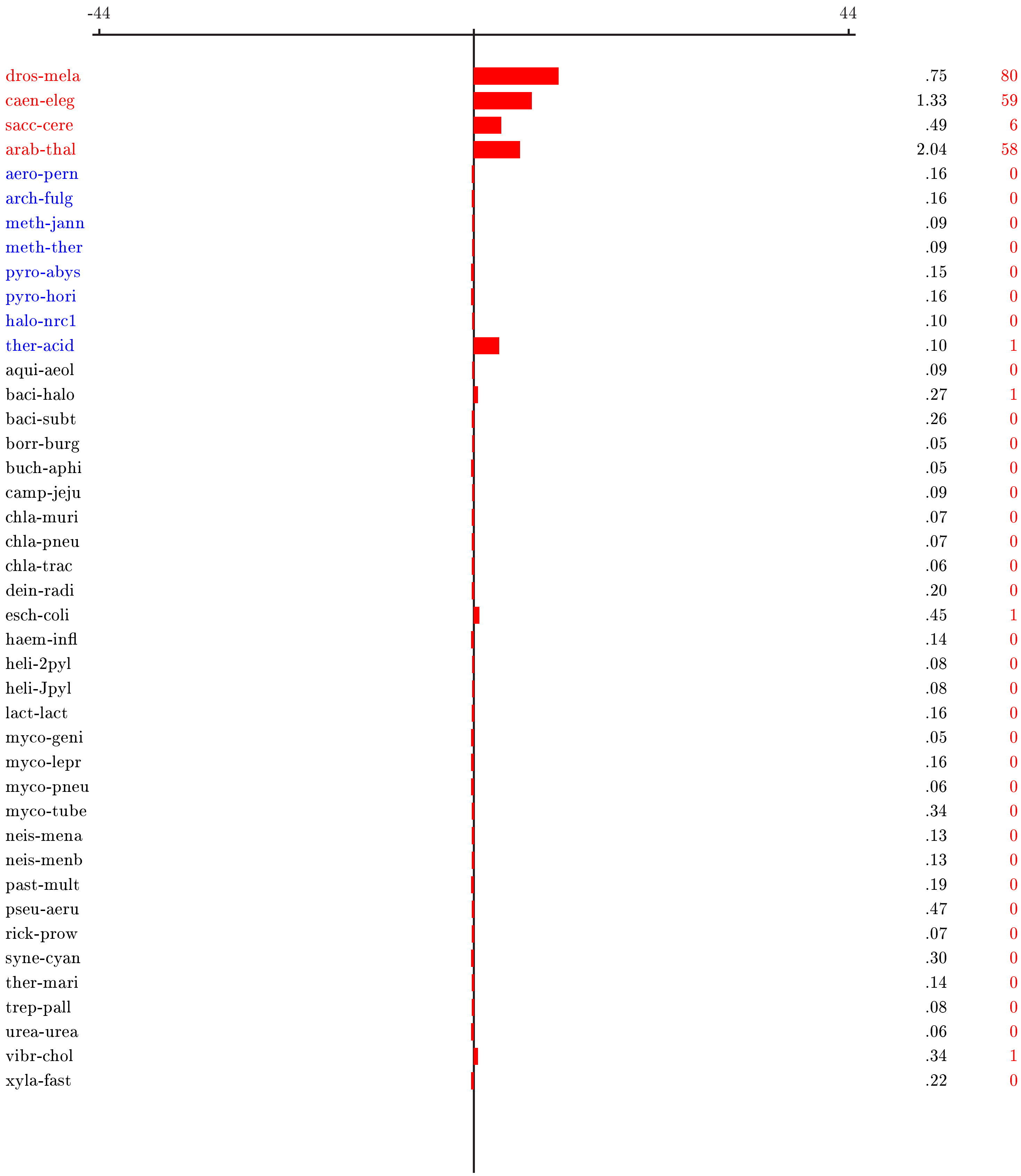
C-x-C-x(5)-G-x(2)-C




| | | |
|---|---|---|
| dros-mela | 2.71 | 280 |
| caen-eleg | 3.67 | 430 |
| sacc-cere | .32 | 4 |
| arab-thal | 4.48 | 13 |
| aero-pern | .05 | 0 |
| arch-fulg | .08 | 0 |
| meth-jann | .07 | 0 |
| meth-ther | .07 | 1 |
| pyro-abys | .01 | 0 |
| pyro-hori | .01 | 0 |
| halo-nrc1 | .02 | 0 |
| ther-acid | .01 | 0 |
| aqui-aeol | .02 | 0 |
| baci-halo | .03 | 2 |
| baci-subt | .04 | 0 |
| borr-burg | .01 | 0 |
| buch-aphi | .02 | 0 |
| camp-jeju | .05 | 0 |
| chla-muri | .09 | 2 |
| chla-pneu | .09 | 0 |
| chla-trac | .09 | 1 |
| dein-radi | .03 | 0 |
| esch-coli | .16 | 0 |
| haem-infl | .04 | 0 |
| heli-2pyl | .04 | 0 |
| heli-Jpyl | .04 | 0 |
| lact-lact | .00 | 0 |
| myco-geni | .00 | 0 |
| myco-lepr | .04 | 0 |
| myco-pneu | .01 | 0 |
| myco-tube | .09 | 1 |
| neis-mena | .05 | 0 |
| neis-menb | .05 | 0 |
| past-mult | .06 | 0 |
| pseu-aeru | .16 | 0 |
| rick-prow | .02 | 0 |
| syne-cyan | .08 | 0 |
| ther-mari | .01 | 0 |
| trep-pall | .17 | 2 |
| urea-urea | .00 | 0 |
| vibr-chol | .09 | 0 |
| xyla-fast | .10 | 0 |

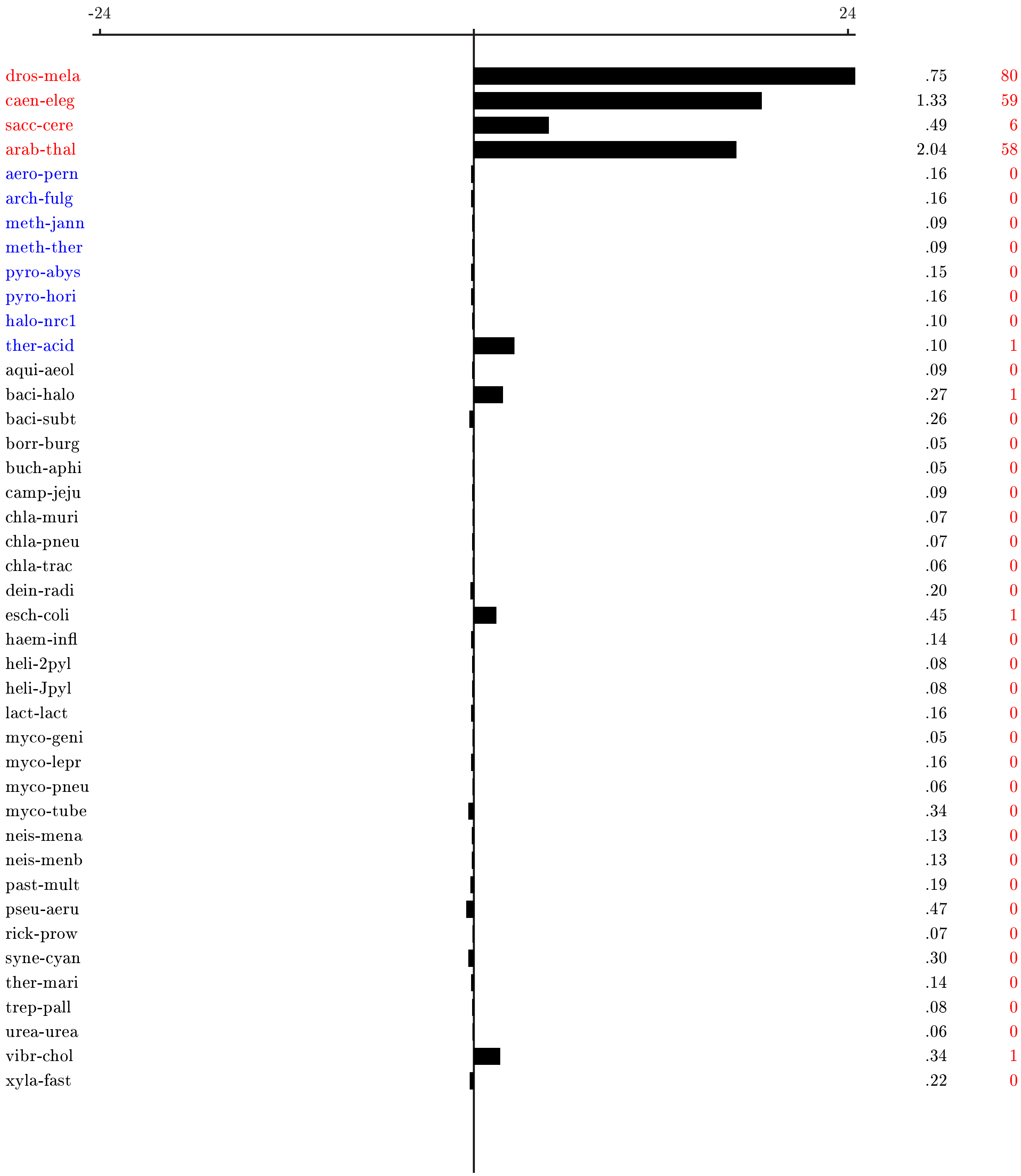
PS00027 - HOMEOBOX-1;PATTERN. 'Homeobox' domain signature.

[LIVMFYG]-[ASLVR]-x(2)-[LIVSTACN]-x-[LIVM]-x(4)-[LIV]-[RKNQESTAIY]-[LIVFSTNKH]-W-[FYVC]-x-[NDQTAH]-x(5)-[RKNAIMW]

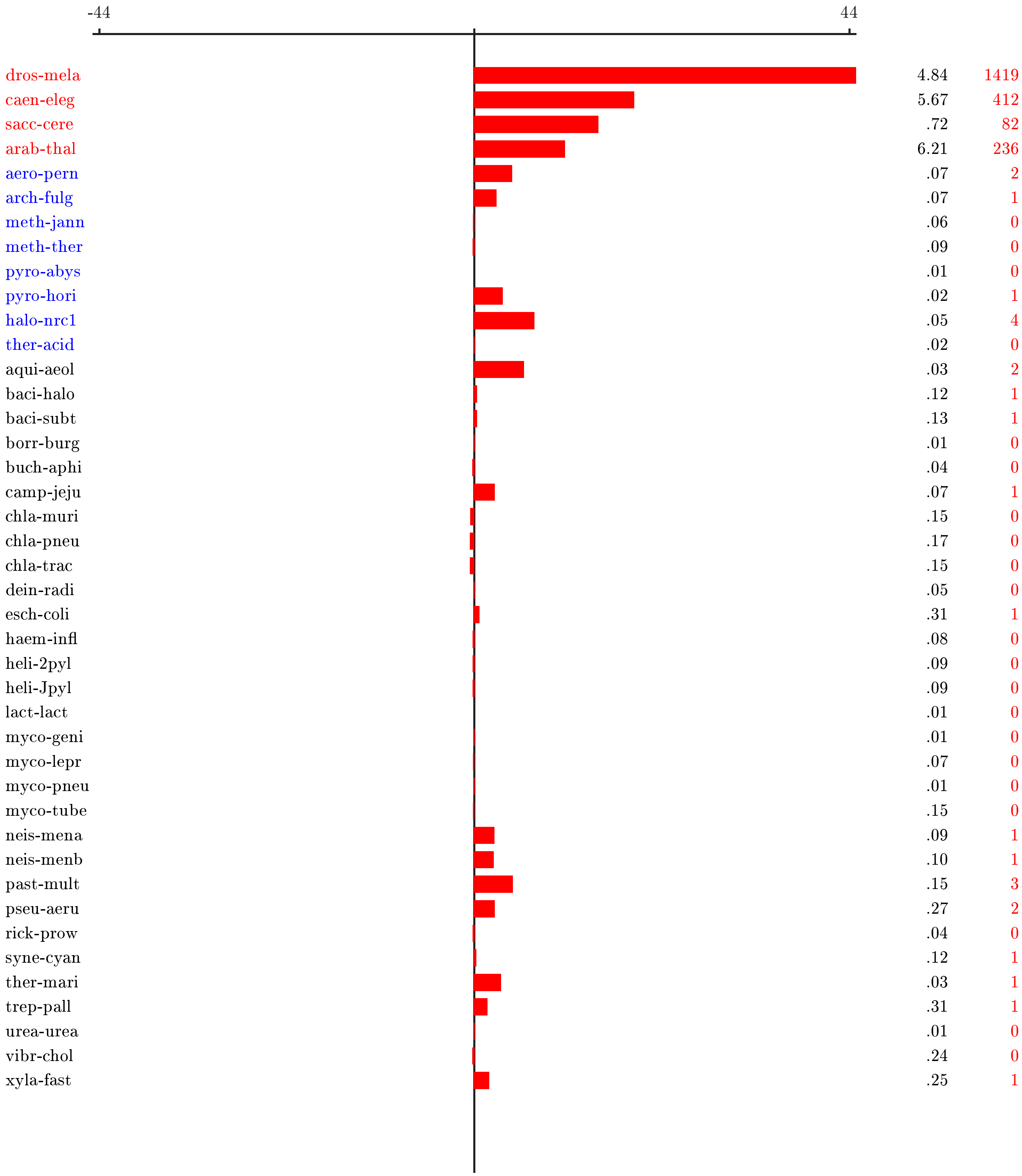


PS00027 HOMEOBOX-1;PATTERN. 'Homeobox' domain signature.

[LIVMFYG]-[ASLVR]-x(2)-[LIVSTACN]-x-[LIVM]-x(4)-[LIV]-[RKNQESTAIY]-[LIVFSTNKH]-W-[FYVC]-x-[NDQTAH]-x(5)-[RKNAIMW]

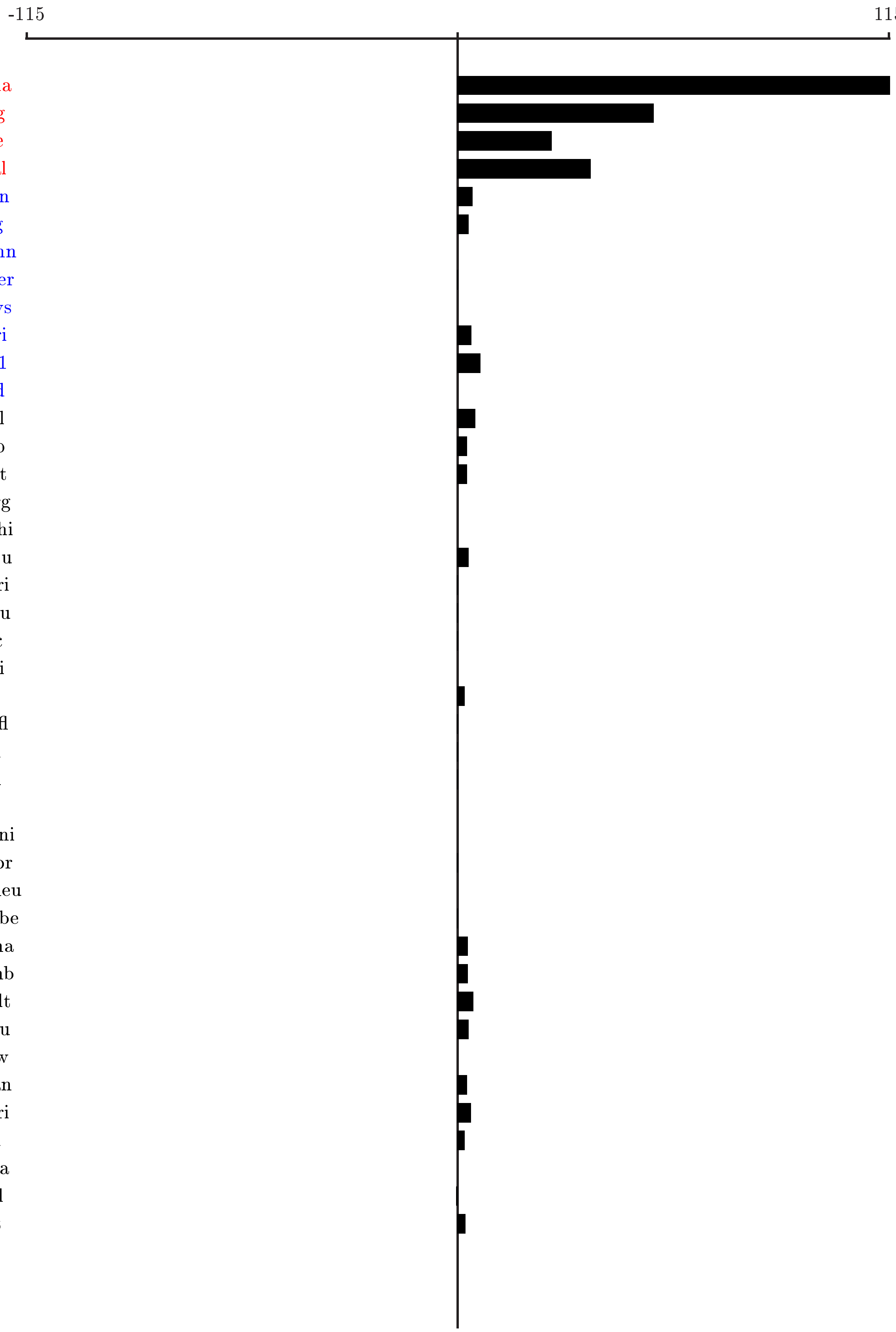


PS00028 - ZINC-FINGER-C2H2;PATTERN. Zinc finger, C2H2 type, domain.
C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H



PS00028 ZINC-FINGER-C2H2;PATTERN. Zinc finger, C2H2 type, domain.

C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H




| | | |
|---|---|---|
| dros-mela | 4.84 | 1419 |
| caen-eleg | 5.67 | 412 |
| sacc-cere | .72 | 82 |
| arab-thal | 6.21 | 236 |
| aero-pern | .07 | 2 |
| arch-fulg | .07 | 1 |
| meth-jann | .06 | 0 |
| meth-ther | .09 | 0 |
| pyro-abys | .01 | 0 |
| pyro-hori | .02 | 1 |
| halo-nrc1 | .05 | 4 |
| ther-acid | .02 | 0 |
| aqui-aeol | .03 | 2 |
| baci-halo | .12 | 1 |
| baci-subt | .13 | 1 |
| borr-burg | .01 | 0 |
| buch-aphi | .04 | 0 |
| camp-jeju | .07 | 1 |
| chla-muri | .15 | 0 |
| chla-pneu | .17 | 0 |
| chla-trac | .15 | 0 |
| dein-radi | .05 | 0 |
| esch-coli | .31 | 1 |
| haem-infl | .08 | 0 |
| heli-2pyl | .09 | 0 |
| heli-Jpyl | .09 | 0 |
| lact-lact | .01 | 0 |
| myco-geni | .01 | 0 |
| myco-lepr | .07 | 0 |
| myco-pneu | .01 | 0 |
| myco-tube | .15 | 0 |
| neis-mena | .09 | 1 |
| neis-menb | .10 | 1 |
| past-mult | .15 | 3 |
| pseu-aeru | .27 | 2 |
| rick-prow | .04 | 0 |
| syne-cyan | .12 | 1 |
| ther-mari | .03 | 1 |
| trep-pall | .31 | 1 |
| urea-urea | .01 | 0 |
| vibr-chol | .24 | 0 |
| xyla-fast | .25 | 1 |

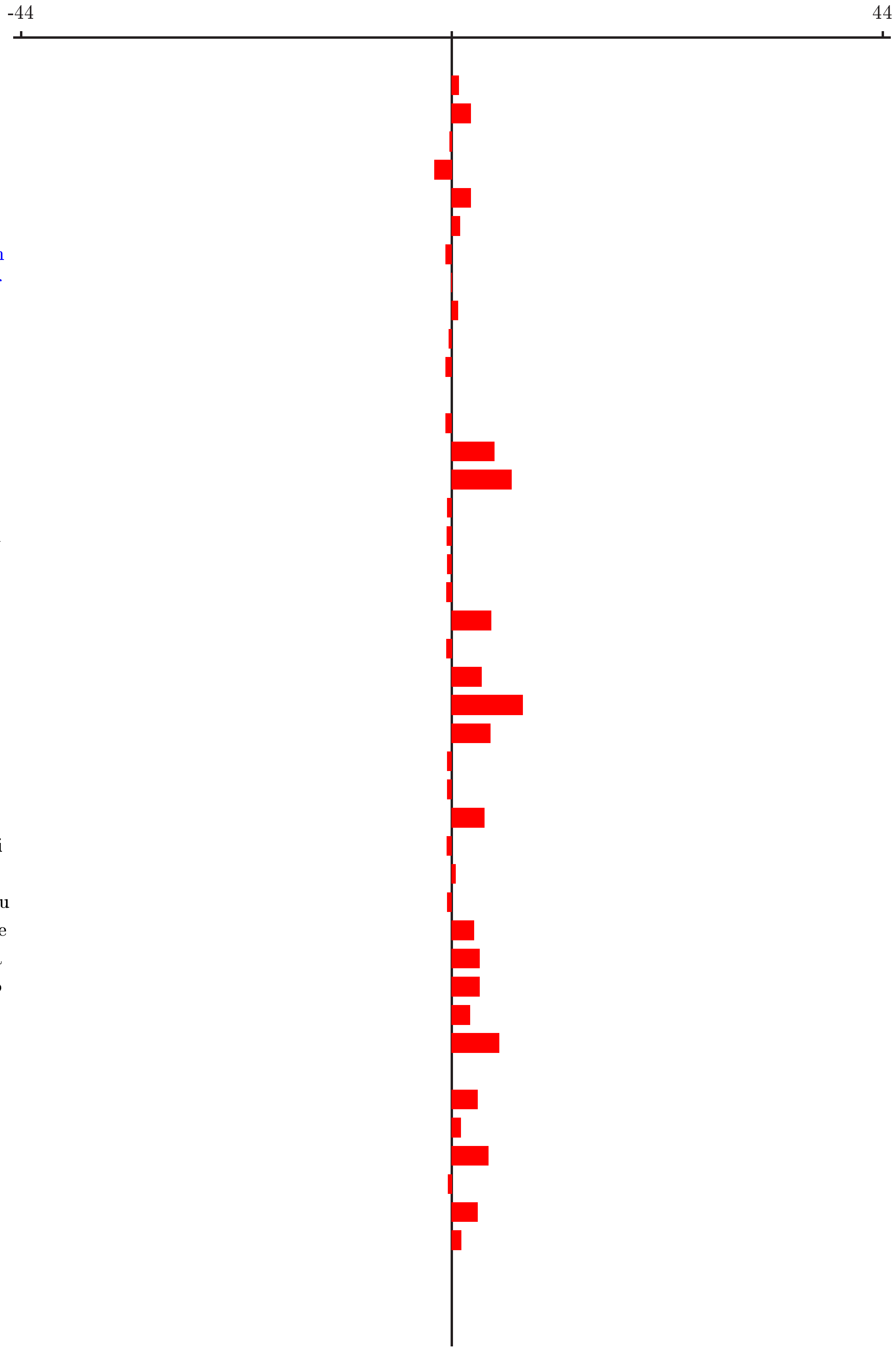
PS00039 - DEAD-ATP-HELICASE;PATTERN. AD-box subfamily ATP-dependent helicases signature.

[LIVMF](2)-D-E-A-D-[RKEN]-x-[LIVMFYGSTN]


| | | |
|---|---|---|
| dros-mela | .68 | 23 |
| caen-eleg | .92 | 23 |
| sacc-cere | .40 | 20 |
| arab-thal | 1.47 | 28 |
| aero-pern | .06 | 1 |
| arch-fulg | .16 | 2 |
| meth-jann | .13 | 1 |
| meth-ther | .14 | 2 |
| pyro-abys | .11 | 0 |
| pyro-hori | .09 | 1 |
| halo-nrc1 | .28 | 0 |
| ther-acid | .09 | 2 |
| aqui-aeol | .09 | 0 |
| baci-halo | .22 | 3 |
| baci-subt | .22 | 3 |
| borr-burg | .06 | 0 |
| buch-aphi | .02 | 1 |
| camp-jeju | .11 | 0 |
| chla-muri | .04 | 0 |
| chla-pneu | .04 | 0 |
| chla-trac | .04 | 0 |
| dein-radi | .12 | 1 |
| esch-coli | .20 | 6 |
| haem-infl | .09 | 3 |
| heli-2pyl | .08 | 0 |
| heli-Jpyl | .08 | 0 |
| lact-lact | .13 | 1 |
| myco-geni | .02 | 0 |
| myco-lepr | .09 | 2 |
| myco-pneu | .03 | 0 |
| myco-tube | .18 | 4 |
| neis-mena | .10 | 2 |
| neis-menb | .10 | 2 |
| past-mult | .10 | 3 |
| pseu-aeru | .32 | 6 |
| rick-prow | .04 | 0 |
| syne-cyan | .13 | 2 |
| ther-mari | .12 | 0 |
| trep-pall | .04 | 1 |
| urea-urea | .04 | 0 |
| vibr-chol | .17 | 10 |
| xyla-fast | .09 | 3 |

PS00043 - HTH-GNTR-FAMILY;PATTERN. Bacterial regulatory proteins, gntR family signature.

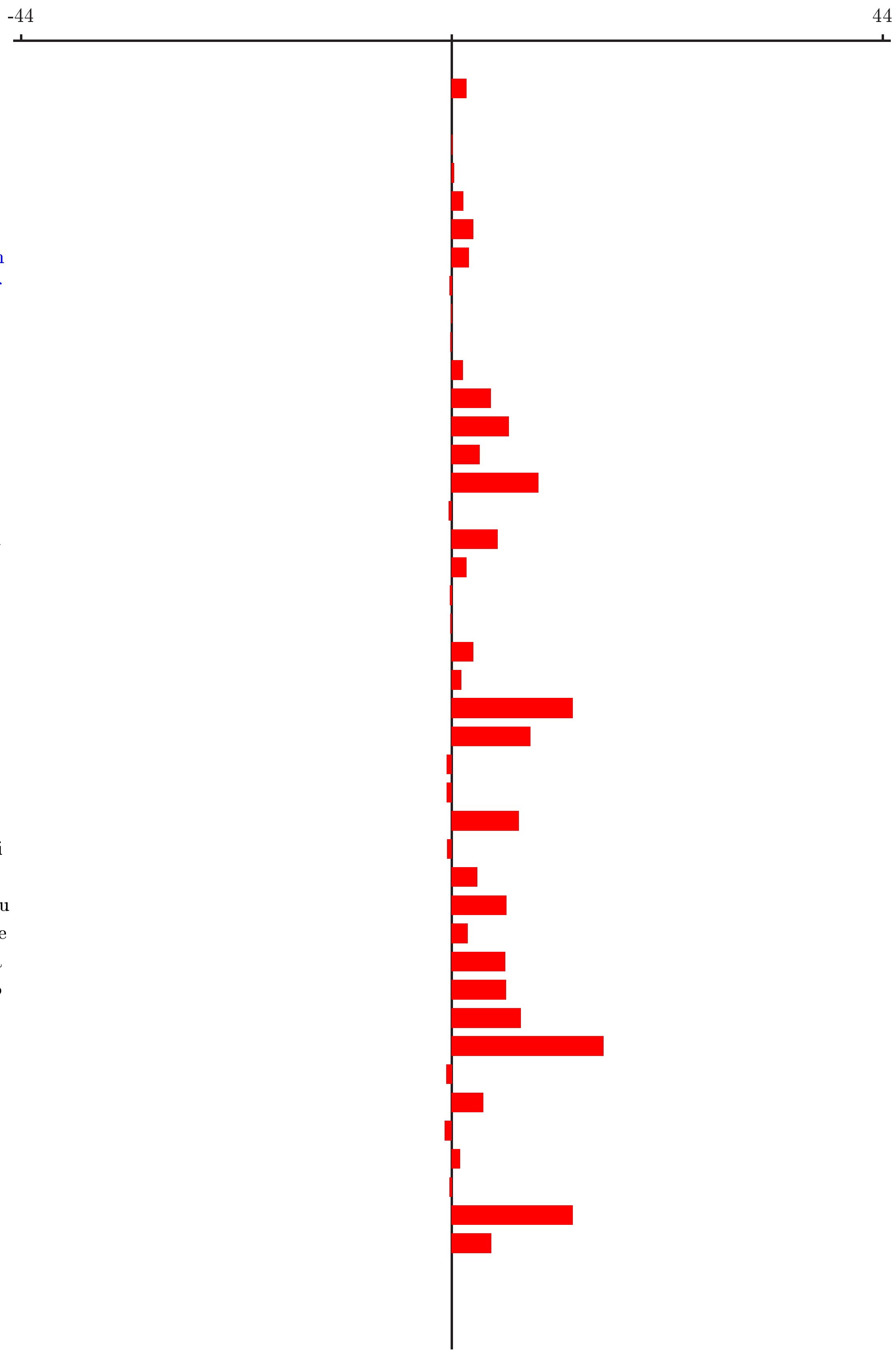
[LIVAPKR]-[PILV]-x-[EQTIVMR]-x(2)-[LIVM]-x(3)-[LIVMFYK]-x-[LIVFT]-[DNGSTK]-[RGTLV]-x-[STAIVP]-[LIVA]-x(2)-[STAGV]-[LIVMFYH]-x(2)-[LMA]




| | | |
|---|---|---|
| dros-mela | 1.41 | 2 |
| caen-eleg | 2.07 | 9 |
| sacc-cere | .74 | 0 |
| arab-thal | 3.37 | 2 |
| aero-pern | .72 | 2 |
| arch-fulg | .54 | 1 |
| meth-jann | .31 | 0 |
| meth-ther | .39 | 0 |
| pyro-abys | .52 | 1 |
| pyro-hori | .55 | 0 |
| halo-nrc1 | .40 | 0 |
| ther-acid | .32 | 0 |
| aqui-aeol | .33 | 0 |
| baci-halo | .77 | 8 |
| baci-subt | .71 | 11 |
| borr-burg | .16 | 0 |
| buch-aphi | .09 | 0 |
| camp-jeju | .23 | 0 |
| chla-muri | .19 | 0 |
| chla-pneu | .21 | 2 |
| chla-trac | .19 | 0 |
| dein-radi | .74 | 4 |
| esch-coli | .95 | 17 |
| haem-infl | .32 | 3 |
| heli-2pyl | .22 | 0 |
| heli-Jpyl | .22 | 0 |
| lact-lact | .41 | 3 |
| myco-geni | .08 | 0 |
| myco-lepr | .56 | 1 |
| myco-pneu | .10 | 0 |
| myco-tube | 1.14 | 4 |
| neis-mena | .33 | 2 |
| neis-menb | .31 | 2 |
| past-mult | .45 | 2 |
| pseu-aeru | 1.29 | 13 |
| rick-prow | .20 | 0 |
| syne-cyan | .67 | 4 |
| ther-mari | .45 | 1 |
| trep-pall | .25 | 2 |
| urea-urea | .09 | 0 |
| vibr-chol | .76 | 4 |
| xyla-fast | .54 | 1 |

PS00044 - HTH-LYSR-FAMILY;PATTERN. Bacterial regulatory proteins, lysR family signature.

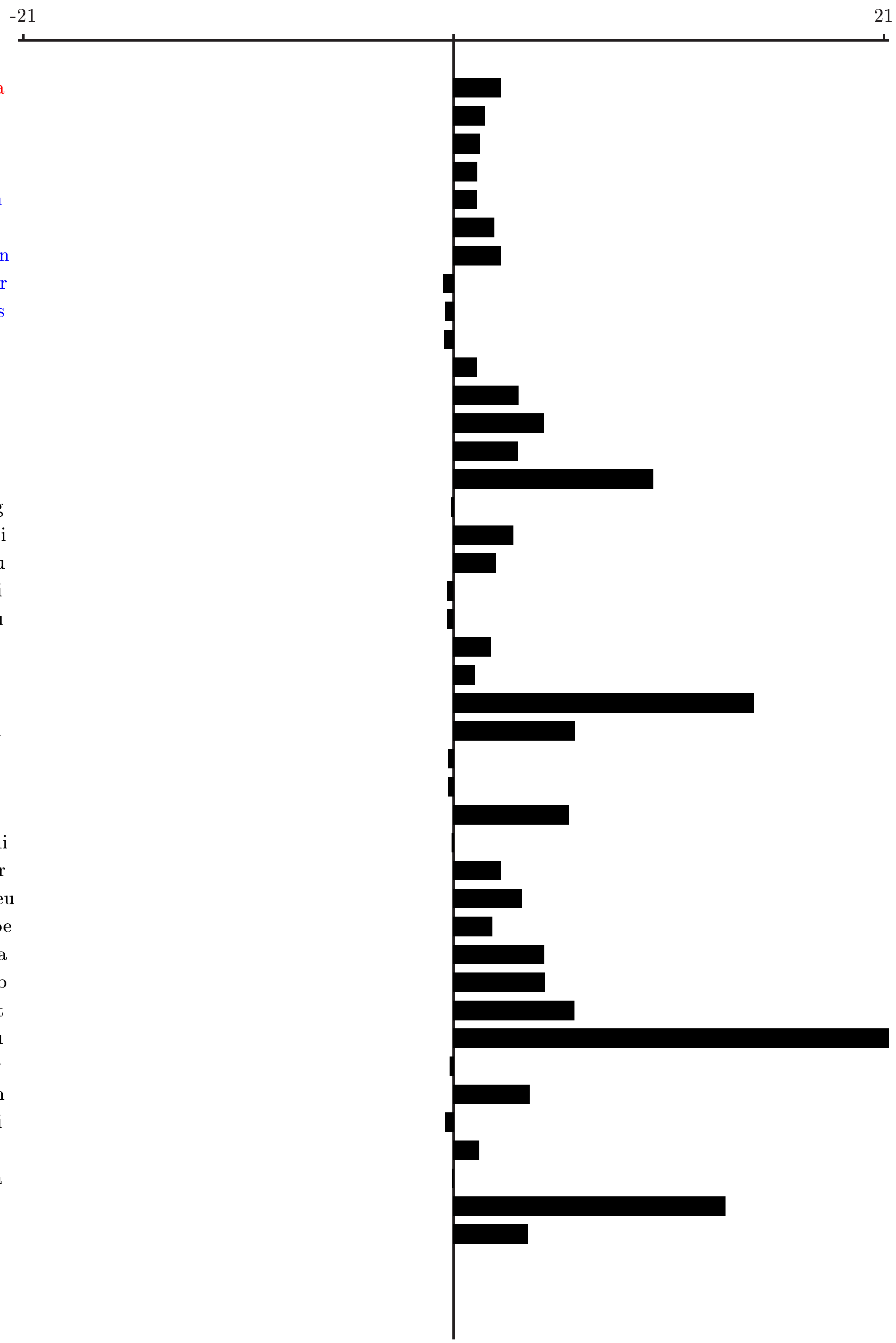
[NQKRHSTAG]-[LIVMFYTA]-x(2)-[STAGLV]-[STAG]-x(4)-[LIVMYCTQR]-[PSTANLVER]-x-[PSTAGQV]-[PSTAGNVMF]-[LIVMFA]-[STAGH]-x(2)-[LIVMF]-x(2)-[LIVMFW]-[RKEAV]-x(2)-[LIVMFYNTAE]-x(3)-[LIMVT]




| | | |
|---|---|---|
| dros-mela | 2.96 | 7 |
| caen-eleg | 3.50 | 6 |
| sacc-cere | 1.10 | 2 |
| arab-thal | 5.69 | 8 |
| aero-pern | 1.23 | 2 |
| arch-fulg | .61 | 2 |
| meth-jann | .16 | 1 |
| meth-ther | .49 | 0 |
| pyro-abys | .40 | 0 |
| pyro-hori | .45 | 0 |
| halo-nrc1 | 1.23 | 2 |
| ther-acid | .40 | 3 |
| aqui-aeol | .24 | 4 |
| baci-halo | 1.07 | 5 |
| baci-subt | .99 | 21 |
| borr-burg | .10 | 0 |
| buch-aphi | .06 | 1 |
| camp-jeju | .22 | 1 |
| chla-muri | .28 | 0 |
| chla-pneu | .28 | 0 |
| chla-trac | .28 | 1 |
| dein-radi | 2.04 | 3 |
| esch-coli | 1.73 | 45 |
| haem-infl | .46 | 8 |
| heli-2pyl | .25 | 0 |
| heli-Jpyl | .25 | 0 |
| lact-lact | .57 | 8 |
| myco-geni | .08 | 0 |
| myco-lepr | 1.33 | 4 |
| myco-pneu | .14 | 2 |
| myco-tube | 3.57 | 7 |
| neis-mena | .64 | 6 |
| neis-menb | .62 | 6 |
| past-mult | .65 | 9 |
| pseu-aeru | 2.81 | 89 |
| rick-prow | .16 | 0 |
| syne-cyan | 1.05 | 6 |
| ther-mari | .40 | 0 |
| trep-pall | .55 | 1 |
| urea-urea | .06 | 0 |
| vibr-chol | 1.33 | 36 |
| xyla-fast | 1.11 | 6 |

PS00044 HTH-LYSR-FAMILY;PATTERN. Bacterial regulatory proteins, lysR family signature.

[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]




| | | |
|---|---|---|
| dros-mela | 2.96 | 7 |
| caen-eleg | 3.50 | 6 |
| sacc-cere | 1.10 | 2 |
| arab-thal | 5.69 | 8 |
| aero-pern | 1.23 | 2 |
| arch-fulg | .61 | 2 |
| meth-jann | .16 | 1 |
| meth-ther | .49 | 0 |
| pyro-abys | .40 | 0 |
| pyro-hori | .45 | 0 |
| halo-nrc1 | 1.23 | 2 |
| ther-acid | .40 | 3 |
| aqui-aeol | .24 | 4 |
| baci-halo | 1.07 | 5 |
| baci-subt | .99 | 21 |
| borr-burg | .10 | 0 |
| buch-aphi | .06 | 1 |
| camp-jeju | .22 | 1 |
| chla-muri | .28 | 0 |
| chla-pneu | .28 | 0 |
| chla-trac | .28 | 1 |
| dein-radi | 2.04 | 3 |
| esch-coli | 1.73 | 45 |
| haem-infl | .46 | 8 |
| heli-2pyl | .25 | 0 |
| heli-Jpyl | .25 | 0 |
| lact-lact | .57 | 8 |
| myco-geni | .08 | 0 |
| myco-lepr | 1.33 | 4 |
| myco-pneu | .14 | 2 |
| myco-tube | 3.57 | 7 |
| neis-mena | .64 | 6 |
| neis-menb | .62 | 6 |
| past-mult | .65 | 9 |
| pseu-aeru | 2.81 | 89 |
| rick-prow | .16 | 0 |
| syne-cyan | 1.05 | 6 |
| ther-mari | .40 | 0 |
| trep-pall | .55 | 1 |
| urea-urea | .06 | 0 |
| vibr-chol | 1.33 | 36 |
| xyla-fast | 1.11 | 6 |

# References

Motifs Statistics

P. N., B. Salvy and P. Flajolet,

ESA 1999, Prague,

to appear in Theoretical Computer Science

Fast Approximate Motif Statistics

P. N.

Journal of Computational Biology, 2001,

Vol. 8, Num. 3